

Article

# Gráfico de control $T^2$ Hotelling para variables cualitativas

Wilson Rojas-Preciado<sup>1,2</sup>, Mauricio J. Rojas-Campuzano<sup>3</sup>, Purificación Galindo-Villardón<sup>2</sup>, Omar Ruiz-Barzola<sup>3</sup>,,,

\* Correspondence: [wrojas@utmachala.edu.ec](mailto:wrojas@utmachala.edu.ec); Tel.: +593-992-83-3719

† Current address: Updated affiliation

‡ These authors contributed equally to this work.

Version December 6, 2021 submitted to Water



<sup>1</sup> **Simple Summary:** A Simple summary goes here.

<sup>2</sup> **Abstract:** La literatura científica es abundante en lo referente a gráficos de control en entornos multivariantes para datos numéricos y mixtos, sin embargo, para datos cualitativos hay pocas publicaciones. Las variables cualitativas aportan valiosa información de procesos en diversos contextos industriales, productivos, sociales. Los procesos educativos no son una excepción, tienen múltiples variables asociadas a estudiantes, profesores e instituciones. Cuando hay muchas variables se corre el riesgo de tomar información redundante o excesiva, luego, es viable la aplicación de métodos multivariantes de reducción de dimensiones para quedarse con pocas variables ficticias, combinación de las reales, que sinteticen la mayor parte de la información. En este contexto se presenta el gráfico de control T2Qv, una técnica de control estadístico de procesos multivariantes que realiza un análisis de datos cualitativos mediante Análisis de correspondencias múltiples (MCA), Análisis Factorial Múltiple y el gráfico  $T^2$  de Hotelling. La interpretación de los puntos fuera de control se realiza comparando los gráficos MCA y analizando la distancia  $X^2$  entre las categorías de la tabla consenso y las que representan puntos fuera de control. El análisis de sensibilidad determinó que el gráfico de control T2Qv tiene un buen rendimiento cuando trabaja con altas dimensiones. Para probar la metodología se hizo un análisis con datos simulados y otro con datos reales relacionados con la educación superior. Para facilitar la difusión y aplicación de la propuesta, se desarrolló un paquete computacional reproducible en R, denominado T2Qv y disponible en GitHub.

<sup>19</sup> **Keywords:** keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.).

## <sup>21</sup> 1. Introduction

<sup>22</sup> Los gráficos de control constituyen una de las herramientas más importantes para definir límites y parámetros óptimos de los procesos, así como para controlar la calidad de los productos mediante la reducción de la variabilidad. El uso de gráficos de control facilita la evaluación del comportamiento de las variables del proceso y contribuye al logro de los objetivos planificados.

<sup>26</sup> La variación de los procesos se entiende como la diversidad de resultados que genera un grupo de variables de un proceso, su monitoreo es un objetivo clave del control estadístico, por lo tanto, es necesario entender los tipos y motivos de la variabilidad. Para ello es preciso registrar de manera sistemática y adecuada diferentes variables del proceso que se desea controlar, como las propiedades de los insumos, las condiciones de operación de los equipos, las competencias del personal que

31 maneja los procesos, además de las características de los productos, la satisfacción de los usuarios, el  
32 cumplimiento de requisitos, entre otras.

33 El pionero del control estadístico de procesos fue Walter Shewhart (SPC). Estableció las diferencias  
34 entre la variabilidad natural o común, presente en todos los procesos, y la provocada por causas  
35 asignables o especiales, que pueden llevarlos a un estado de fuera de control. Señaló que un proceso  
36 está en control estadístico cuando trabaja sólo con causas comunes de variación. Propuso los primeros  
37 gráficos de control para variables de tipo continuo y para variables de atributos [1].

38 El SPC mediante gráficos de control permitió a las organizaciones monitorear el comportamiento  
39 de una variable a la vez, no obstante, las organizaciones requirieron, con el pasar del tiempo, el  
40 análisis de varias características de calidad de forma simultánea, abriendo la puerta al SPC desde  
41 una perspectiva multivariante [2]. Para facilitar el control de calidad de procesos es común el uso de  
42 gráficas de control que recolectan abundante información en diversas variables de forma simultánea,  
43 su análisis permite caracterizar los diferentes tipos de variables que afectan la calidad y explican su  
44 comportamiento a lo largo del tiempo [3].

45 Hay una variedad de gráficos de control de procesos desde la perspectiva multivariante, entre  
46 los clásicos están el Gráfico  $T^2$  de Hotelling [4], el Multivariate Exponentially Weighted Moving –  
47 MEWMA [5], el Multivariate Cumulative Sum Control Chart – MCUSUM [6]. Con el transcurso  
48 del tiempo se hicieron diversos aportes para mejorar el rendimiento de estos gráficos, entre los más  
49 destacados están Gráfico de control  $T^2$  con tamaños de muestra adaptables [7], Gráfico de control  
50  $T^2$  con intervalos de muestreo variables [8], Gráfico de control  $T^2$  con líneas de advertencia dobles  
51 [9], Gráfico de control robusto [10], Gráficos de control basados en modelos de minería de datos para  
52 procesos multivariantes y autocorrelacionados [11], Gráficos de control de calidad multivariantes con  
53 dimensión variable [12], Gráfico de control para el coeficiente de variación multivariante [13].

54 Además de estos gráficos de control para entornos paramétricos, se desarrollaron otros para  
55 datos numéricos y cualitativos en entornos multivariantes no paramétricos, entre ellos el Gráfico de  
56 control multivariante basado en la distancia de Gower para una combinación de variables continuas y  
57 cualitativas [14], Gráfico de control multivariante basado en la combinación de PCA para características  
58 de calidad de atributos y variables [15], Gráfico de control multivariante no paramétrico basado en la  
59 ponderación de novedad sensible a la densidad para procesos no normales [16], Gráfico de control  
60 de deméritos con clustering difuso de c-medias [17], Gráfico de control basado en ACP que utiliza  
61 máquinas de vectores de soporte para distribuciones no normales multivariadas [18], Gráfico CUSUM  
62 no paramétrico para monitorear procesos multivariados correlacionados en serie [19], Gráfico de  
63 control multivariante basado en Kernel PCA para monitorear características de calidad de atributos y  
64 variables mixtas [20], Gráfico  $T^2$  basado en la combinación de PCA para datos continuos y cualitativos  
65 con detección de datos atípicos [21].

66 Como se puede observar, la literatura científica es abundante en lo referente a gráficos de control  
67 en entornos multivariantes paramétricos y no paramétricos para datos numéricos y, en los últimos años,  
68 para datos mixtos (numéricos y cualitativos), sin embargo, son pocas las publicaciones sobre gráficos  
69 de control multivariantes para datos cualitativos. En este campo las propuestas se han desarrollado  
70 alrededor del análisis de variables que siguen una distribución Poisson y el análisis de variables  
71 multinomiales.

72 La primera propuesta fue la de Holgate [22], quien presentó un trabajo sobre la distribución  
73 Poisson bivariante para variables correlacionadas. Este modelo fue tomado como insumo en las  
74 investigaciones de autores como Chiu and Kuo [23], Lee and Costa [24], Laungrungrong *et al.* [25],  
75 Epprecht *et al.* [26]. Otra propuesta destacada es la de Lu [27], quien desarrolló un gráfico de control  
76 tipo Shewhart para procesos multivariados con variables de atributos, cuando la característica de  
77 calidad asume valores binarios, que se denominó gráfico  $np$  multivariante (MNP). No obstante, hay  
78 escenarios en los que una clasificación dicotómica es insuficiente y se vuelve necesario acudir a niveles  
79 intermedios, en cuyo caso el análisis requiere el uso de distribuciones multinomiales.

80 En este contexto Mukhopadhyay [28] planteó un gráfico de control multivariante utilizando el  
81 estadístico  $D^2$  de Mahalanobis para atributos que siguen una distribución multinomial. Además,  
82 surgieron los gráficos de control multivariantes en procesos multinomiales bajo el enfoque difuso  
83 [29]; Taleb [30] introdujo gráficos de control para el monitoreo de procesos multivariados con datos  
84 lingüísticos multidimensionales, basados en dos procedimientos: la teoría de la probabilidad y la teoría  
85 difusa; Fernández *et al.* [31] presentaron un gráfico de control multivariante multinomial T2 con un  
86 enfoque difuso.

87 Un aporte interesante es el de Epprecht *et al.* [26], quienes presentaron una combinación lineal  
88 óptima de variables discretas, cuando siguen la distribución de Poisson, para el SPC multivariantes.  
89 Asimismo, Ali and Aslam [32] desarrollaron gráficos de control para datos con distribución Poisson  
90 multivariante utilizando un muestreo generalizado de estados dependientes múltiples (GMDS).

91 En el estudio de los procesos que se desarrollan en el entorno social-educativo y que explican  
92 el comportamiento de variables como el rendimiento académico, tasas de graduación o deserción,  
93 producción científica, porcentajes de matrícula de nuevo ingreso, entre otros, se maneja con mucha  
94 frecuencia variables cualitativas. No es que estén ausentes los datos cuantitativos, sino que, en  
95 las bases de datos que se utilizan para estos análisis, abundan las variables cualitativas nominales  
96 y ordinales sobre las de tipo numérico, algunos ejemplos de datos de los estudiantes son: sexo,  
97 lugar de procedencia, autodenominación étnica, grado académico de los padres, tipo de institución  
98 educativa de procedencia (fiscal, particular, municipal); ejemplos de datos de las instituciones son:  
99 tipo de sostenimiento económico, jornada, modalidad, campo de estudio, niveles (tecnológico, grado  
100 y postgrado), tipo de infraestructura; ejemplos asociados a datos de los profesores son: titularidad,  
101 dedicación, grado académico, grado en el escalafón, discapacidad, entre otros.

102 López [33] señala que al observar muchas variables sobre una muestra es presumible que una  
103 parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los métodos  
104 multivariantes de reducción de la dimensión tratan de eliminarla combinando muchas variables  
105 observadas para quedarse con pocas variables ficticias que, aunque no observadas, sean combinación  
106 de las reales y sinteticen la mayor parte de la información contenida en sus datos. En este caso se  
107 deberá tener en cuenta el tipo de variables que maneja. Si son variables cuantitativas las técnicas que  
108 le permiten este tratamiento pueden ser el Análisis de componentes principales [34,35], el Análisis  
109 factorial [36–38], mientras que, si se trata de variables cualitativas, es recomendable la aplicación de  
110 un Análisis de correspondencias múltiple, Análisis de homogeneidad o un Análisis de Escalamiento  
111 multidimensional.

112 En el contexto del SPC el análisis gráficos de control para variables cualitativas todavía es  
113 incipiente. Al analizar los procedimientos publicados por los autores citados en este estudio, se detecta  
114 limitaciones que podrían restringir su aplicación, por ejemplo, el análisis de pocas características de la  
115 calidad, el uso de muestras constituidas por elementos individuales en vez de grupos, la dificultad de  
116 trabajar con muchas categorías de forma simultánea. Surge, entonces, la necesidad de un gráfico de  
117 control para la representación de p variables cualitativas, que pueda trabajar con múltiples categorías  
118 nominales y ordinales y que facilite la identificación de las causas que pueden llevar al proceso a un  
119 estado fuera de control.

120 Esta necesidad se atiende en esta investigación, cuyo objetivo es desarrollar un gráfico de control  
121 para variables cualitativas con múltiples categorías nominales y ordinales, mediante la aplicación de  
122 una metodología de análisis multivariante, para que se contribuya a la diversificación de técnicas en la  
123 fase I del control estadístico de procesos.

124 Este artículo está organizado de la siguiente manera: la Introducción, que establece los  
125 antecedentes conceptuales y referenciales de los gráficos de control multivariantes aplicados a variables  
126 cualitativas; la sección 2, materiales y métodos, que detalla el procedimiento que se siguió en el  
127 desarrollo del gráfico de control propuesto; la sección 3 describe al complemento computacional que  
128 facilita la aplicación de esta metodología; la sección 4 muestra los resultados mediante el análisis de  
129 datos simulados y datos reales aplicados al contexto de la educación superior; la sección 5 corresponde

130 al análisis de sensibilidad que relaciona el número de dimensiones analizadas versus la confiabilidad  
 131 de los resultados. La sección 6 presenta la discusión mediante un análisis comparativo entre el gráfico  
 132 de control T2Qv y las propuestas de otros autores. Finalmente, la sección 7 establece las conclusiones.

133 **2. Metodología**

134 **2.1. Notación**

135 La tabla 1 contiene elementos, representación y ejemplos de la manera como se presentan los  
 136 elementos algebraicos abordados en la metodología.

| Elementos                              | Representación                       | Ejemplo                  |
|--|--------------------------------------|--------------------------|
| Escalares                              | Letras en minúscula.                 | $v, \lambda$             |
| Vectores                               | Letras en minúscula y en negrita.    | $\mathbf{v}, \mathbf{u}$ |
| Matrices                               | Letras en mayúscula y en negrita.    | $\mathbf{V}, \mathbf{X}$ |
| Matrices de tres vías (Cubos de datos) | Letras con doble trazo en mayúscula. | $\mathbb{C}, \mathbb{X}$ |

Table 1. Elementos algebraicos

137 A lo largo del artículo se utilizarán letras para hacer referencia a parámetros necesarios, se los  
 138 enuncia a continuación en la tabla 2:

| Letra | Significado  | Especificación     |
|-------|--|--------------------|
| $p$   | Número de dimensiones  |                    |
| $K$   | Número total de tablas (Especifica la profundidad del cubo de datos) |                    |
| $k$   | Índice de tabla  | $k=1, 2, \dots, K$ |
| $T$   | Índice de matriz transpuesta   | $\mathbf{x}^T$     |
| $n$   | Tamaño muestral de las $k$ tablas                                    |                    |

Table 2. Notación

139 **2.2. Bases metodológicas**

140 **2.2.1. Análisis de Correspondencias Simple**

141 El tratamiento multivariante de variables cualitativas requiere un proceso metodológico distinto,  
 142 uno de los más representativos es el Análisis de Correspondencias [39]. Según [33], este análisis implica  
 143 estudios de similaridad o disimilaridad entre categorías, se debe cuantificar la diferencia o distancia  
 144 entre ellas sumando las diferencias cuadráticas relativas entre las frecuencias de las distribuciones  
 145 de las variables analizadas, lo que conduce al concepto de la  $\chi^2$ . Así, el análisis de correspondencias  
 146 puede considerarse como un análisis de componentes principales aplicado a variables cualitativas que,  
 147 al no poder utilizar correlaciones, se basa en la distancia no euclídea de la  $\chi^2$ .

148 En el enfoque francés del análisis de correspondencias, que se caracteriza por el énfasis en la  
 149 geometría, el análisis de una tabla cruzada se llama análisis de correspondencias (CA) y el análisis de  
 150 una colección de matrices indicadoras, se denomina análisis de correspondencias múltiples (MCA)  
 151 [40]. En contextos anglosajones, el MCA es conocido como Análisis de Homogeneidad o Escalamiento  
 152 Dual, especialmente en psicometría.

153 **2.2.2. Análisis de Correspondencias Múltiples (MCA)**

154 El análisis de correspondencias múltiples (MCA) es una generalización del análisis de  
 155 correspondencias simple o binario, donde se incluyen más variables cualitativas. Se obtiene al realizar

<sup>156</sup> el análisis de correspondencias simple a una tabla disyuntiva completa, conocida como la tabla de  
<sup>157</sup> Burt.

<sup>158</sup> Se tiene una matriz de datos con  $p$  variables cualitativas, cada una con  $h$  categorías ( $h$   
<sup>159</sup> >1). En el ejemplo que se desarrolla para esta investigación, se dispone de una base de datos  
<sup>160</sup> (*Datak10Contaminated*) constituida por 10 tablas, cada una tiene 10 variables y cada variable, 3 categorías  
<sup>161</sup> (Alto, Medio y Bajo).

|       | $V_1$ | $V_2$   | $\dots$ | $V_p$ |
|-------|-------|---------|---------|-------|
| Alto  | Medio | $\dots$ | Medio   |       |
| Medio | Bajo  | $\dots$ | Alto    |       |
| :     | :     | :       | :       |       |
| Bajo  | Alto  | $\dots$ | Bajo    |       |

**Table 3.** Matriz inicial

<sup>162</sup> Esta matriz es equivalente a la matriz disyuntiva  $Z$ , que desglosa las variables en cada una de sus  
<sup>163</sup> modalidades y se registra la ocurrencia de eventos de forma binaria.

| $V_1$<br>Alto | $V_1$<br>Medio | $V_1$<br>Bajo | $V_2$<br>Alto | $V_2$<br>Medio | $V_2$<br>Bajo | $\dots$ | $V_p$<br>Alto | $V_p$<br>Medio | $V_p$<br>Bajo |
|---------------|----------------|---------------|---------------|----------------|---------------|---------|---------------|----------------|---------------|
| 1             | 0              | 0             | 1             | 0              | 0             | ...     | 0             | 1              | 0             |
| 0             | 1              | 0             | 0             | 1              | 0             | ...     | 1             | 0              | 0             |
| 0             | 0              | 1             | 0             | 0              | 1             | ...     | 0             | 0              | 1             |
| :             | :              | :             | :             | :              | :             | ..      | :             | :              | :             |
| 1             | 0              | 0             | 1             | 0              | 0             | ...     | 1             | 0              | 0             |

**Table 4.** Matriz disyuntiva  $Z$

<sup>164</sup> La tabla de Burt viene dada por:

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z} \quad (1)$$

<sup>165</sup> La construcción de la matriz de Burt se da por la superposición de tablas. En las tablas ubicadas  
<sup>166</sup> en la diagonal se encuentran matrices diagonales que contienen las frecuencias marginales de cada  
<sup>167</sup> una de las variables. Fuera de la diagonal de la matriz de Burt se encuentran las tablas cruzadas por  
<sup>168</sup> pares de variables.

<sup>169</sup> Para realizar el análisis de correspondencias múltiples se parte de la matriz de Burt, obtenida con la  
<sup>170</sup> ecuación 1. Esta matriz está formada por las frecuencias absolutas, éstas se transforman en frecuencias  
<sup>171</sup> relativas, dividiendo los valores de la matriz por la frecuencia total, dando lugar a una nueva matriz  
<sup>172</sup> que se denominará  $\mathbf{P}$ .

<sup>173</sup> Se obtienen las marginales de las filas ( $mf$ ) y de las columnas ( $mc$ ) de la matriz  $\mathbf{P}$  (Tabla 5). A estos  
<sup>174</sup> vectores se los conoce también como *Masas de fila y columna*, respectivamente.

<sup>175</sup> Se obtiene la matriz de residuos estandarizados  $\mathbf{S}$ .

$$\mathbf{S} = \mathbf{D}_{\text{fila}}^{-\frac{1}{2}} (\mathbf{P} - \mathbf{mf} \mathbf{mc}') \mathbf{D}_{\text{columna}}^{-\frac{1}{2}} \quad (2)$$

<sup>176</sup> donde:

- <sup>177</sup> •  $\mathbf{D}_{\text{fila}}$  es una matriz diagonal que contiene las masas de las filas.
- <sup>178</sup>
- <sup>179</sup> •  $\mathbf{D}_{\text{columna}}$  es una matriz diagonal que contiene las masas de las columnas

|               | $V_1 : Alto$ | $V_1 : Medio$ | $V_1 : Bajo$ | $V_2 : Alto$ | $V_2 : Medio$ | $V_2 : Bajo$ | $\dots$  | $V_p : Alto$    | $V_p : Medio$   | $V_p : Bajo$ |
|---------------|--------------|---------------|--------------|--------------|---------------|--------------|----------|-----------------|-----------------|--------------|
| $V_1 : Alto$  | $b_{1,1}$    | 0             | 0            | $b_{1,4}$    | $b_{1,5}$     | $b_{1,6}$    | $\dots$  | $b_{1,3p-2}$    | $b_{1,3p-1}$    | $b_{1,3p}$   |
| $V_1 : Medio$ | 0            | $b_{2,2}$     | 0            | $b_{2,4}$    | $b_{2,5}$     | $b_{2,6}$    | $\dots$  | $b_{2,3p-2}$    | $b_{2,3p-1}$    | $b_{2,3p}$   |
| $V_1 : Bajo$  | 0            | 0             | $b_{3,3}$    | $b_{3,4}$    | $b_{3,5}$     | $b_{3,6}$    | $\dots$  | $b_{3,3p-2}$    | $b_{3,3p-1}$    | $b_{3,3p}$   |
| $V_2 : Alto$  | $b_{4,1}$    | $b_{4,2}$     | $b_{4,3}$    | $b_{4,4}$    | 0             | 0            | $\dots$  | $b_{4,3p-2}$    | $b_{4,3p-1}$    | $b_{4,3p}$   |
| $V_2 : Medio$ | $b_{5,1}$    | $b_{5,2}$     | $b_{5,3}$    | 0            | $b_{5,5}$     | 0            | $\dots$  | $b_{5,3p-2}$    | $b_{5,3p-1}$    | $b_{5,3p}$   |
| $V_2 : Bajo$  | $b_{6,1}$    | $b_{6,2}$     | $b_{6,3}$    | 0            | 0             | $b_{6,6}$    | $\dots$  | $b_{6,3p-2}$    | $b_{6,3p-1}$    | $b_{6,3p}$   |
| $\vdots$      | $\vdots$     | $\vdots$      | $\vdots$     | $\vdots$     | $\vdots$      | $\vdots$     | $\ddots$ | $\vdots$        | $\vdots$        | $\vdots$     |
| $V_p : Alto$  | $b_{3p-2,1}$ | $b_{3p-2,2}$  | $b_{3p-2,3}$ | $b_{3p-2,4}$ | $b_{3p-2,5}$  | $b_{3p-2,6}$ | $\dots$  | $b_{3p-2,3p-2}$ | 0               | 0            |
| $V_p : Medio$ | $b_{3p-1,1}$ | $b_{3p-1,2}$  | $b_{3p-1,3}$ | $b_{3p-1,4}$ | $b_{3p-1,5}$  | $b_{3p-1,6}$ | $\dots$  | 0               | $b_{3p-1,3p-1}$ | 0            |
| $V_p : Bajo$  | $b_{3p,1}$   | $b_{3p,2}$    | $b_{3p,3}$   | $b_{3p,4}$   | $b_{3p,5}$    | $b_{3p,6}$   | $\dots$  | 0               | 0               | $b_{3p,3p}$  |

Table 5. P: Tabla de contingencia de Burt en frecuencias relativas

| $V_1 : Alto$    | $V_1 : Medio$   | $V_1 : Bajo$    | $V_2 : Alto$    | $V_2 : Medio$   | $V_2 : Bajo$    | $\dots$ | $V_p : Alto$       | $V_p : Medio$      | $V_p : Bajo$     |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|---------|--------------------|--------------------|------------------|
| $b_{\bullet,1}$ | $b_{\bullet,2}$ | $b_{\bullet,3}$ | $b_{\bullet,4}$ | $b_{\bullet,5}$ | $b_{\bullet,6}$ | $\dots$ | $b_{\bullet,3p-2}$ | $b_{\bullet,3p-1}$ | $b_{\bullet,3p}$ |

Table 6. Frecuencias marginales de las filas. (mf)

| $V_1 : Alto$    | $V_1 : Medio$   | $V_1 : Bajo$    | $V_2 : Alto$    | $V_2 : Medio$   | $V_2 : Bajo$    | $\dots$ | $V_p : Alto$       | $V_p : Medio$      | $V_p : Bajo$     |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|---------|--------------------|--------------------|------------------|
| $b_{\bullet,1}$ | $b_{\bullet,2}$ | $b_{\bullet,3}$ | $b_{\bullet,4}$ | $b_{\bullet,5}$ | $b_{\bullet,6}$ | $\dots$ | $b_{\bullet,3p-2}$ | $b_{\bullet,3p-1}$ | $b_{\bullet,3p}$ |

Table 7. Frecuencias marginales de las columnas. (mc)

180 Se aplica descomposición singular (SVD) a la matriz  $\mathbf{S}$  (Ecuación 2):

$$\mathbf{S} = \mathbf{UDV}' \quad (3)$$

181 donde:

182 •  $\mathbf{U}$  y  $\mathbf{V}$  son matrices ortogonales.

183

184 •  $\mathbf{D}$  es una matriz diagonal que contiene los valores singulares.

185 Para encontrar las coordenadas estandarizadas se aplica lo siguiente:

$$\mathbf{X} = \mathbf{D}_{\text{fila}}^{-\frac{1}{2}} \mathbf{U} \quad (4)$$

$$\mathbf{Y} = \mathbf{D}_{\text{columna}}^{-\frac{1}{2}} \mathbf{V} \quad (5)$$

186 Para los fines necesarios, se utilizará las coordenadas de las columnas (Tabla 8).

### 187 2.2.3. Análisis de Homogeneidad

188 El Análisis de Homogeneidad, Homogeneous Alternating Least Squares (HOMALS), es un  
 189 modelo de la familia de modelos matemáticos del Escalamiento óptimo del sistema Gifi [41], el cual  
 190 comprende una serie de técnicas exploratorias de análisis multivariado no lineal. Igual que el MCA,  
 191 HOMALS se considera una forma de Análisis de Componentes Principales para datos cualitativos. El  
 192 Análisis de Homogeneidad representa los objetos analizados mediante puntos en el modelo espacial,  
 193 sus características más relevantes se presentan en las relaciones geométricas entre los puntos, para ello,

|               | $Dim_1$          | $Dim_2$          | $\dots$  | $Dim_{3p}$       |
|---------------|------------------|------------------|----------|------------------|
| $V_1 : Alto$  | $v_1 d_{1alto}$  | $v_1 d_{1alto}$  | $\dots$  | $v_1 d_{palto}$  |
| $V_1 : Medio$ | $v_1 d_{1medio}$ | $v_1 d_{1medio}$ | $\dots$  | $v_1 d_{pmedio}$ |
| $V_1 : Bajo$  | $v_1 d_{1bajo}$  | $v_1 d_{1bajo}$  | $\dots$  | $v_1 d_{pbajo}$  |
| $\vdots$      | $\vdots$         | $\vdots$         | $\ddots$ | $\vdots$         |
| $V_p : Bajo$  | $v_p d_{1bajo}$  | $v_p d_{1bajo}$  | $\dots$  | $v_p d_{pbajo}$  |

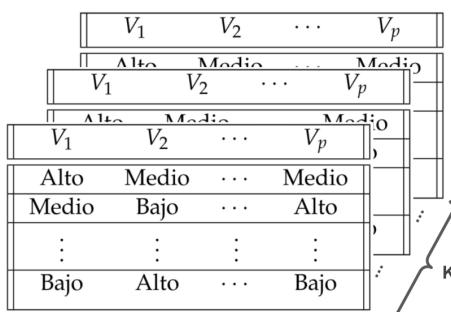
**Table 8.** Coordenadas estandarizadas de las columnas.

<sup>194</sup> es necesario la cuantificación de datos cualitativos [42].

<sup>195</sup> El uso de variables cualitativas no es particularmente restrictivo, ya que una variable numérica continua  
<sup>196</sup> se puede considerar como una variable cualitativa con un gran número de categorías. HOMALS se  
<sup>197</sup> diferencia del el MCA en que éste utiliza la función de Descomposición de valores propios mientras  
<sup>198</sup> que el Análisis de Homogeneidad utiliza Mínimos Cuadrados Alternos, lo que se conoce en la literatura  
<sup>199</sup> como la Solución de Homals [40].

#### <sup>200</sup> 2.2.4. Generalización a $k$ tablas

<sup>201</sup> Si se tienen  $k$  tablas, con la misma estructura de la tabla 3, como se visualiza en la figura 1, se aborda  
<sup>202</sup> el enfoque del análisis factorial múltiple (MFA). Escofier and Pagès [43] indica que el MFA utiliza  
<sup>203</sup> análisis de correspondencias múltiples cuando se trata de variables cualitativas. El procedimiento  
<sup>204</sup> implica la realización de un MCA por cada tabla y dividirlo por su primer valor propio con la finalidad  
<sup>205</sup> de obtener  $K$  grupos normalizados. Posteriormente se consideran todas las tablas y se realiza un MCA  
<sup>206</sup> global.

**Figure 1.**  $k$  tablas con el formato inicial.

<sup>207</sup> La generalización a  $k$  tablas del procedimiento del MCA, se presenta en la Figura 2

<sup>208</sup> Se llama  $C$  a cada tabla de coordenadas. Con la finalidad de detectar la magnitud de las variables  
<sup>209</sup> latentes, su aporte neto a las variables, se trata la matriz  $C$  con valor absoluto.

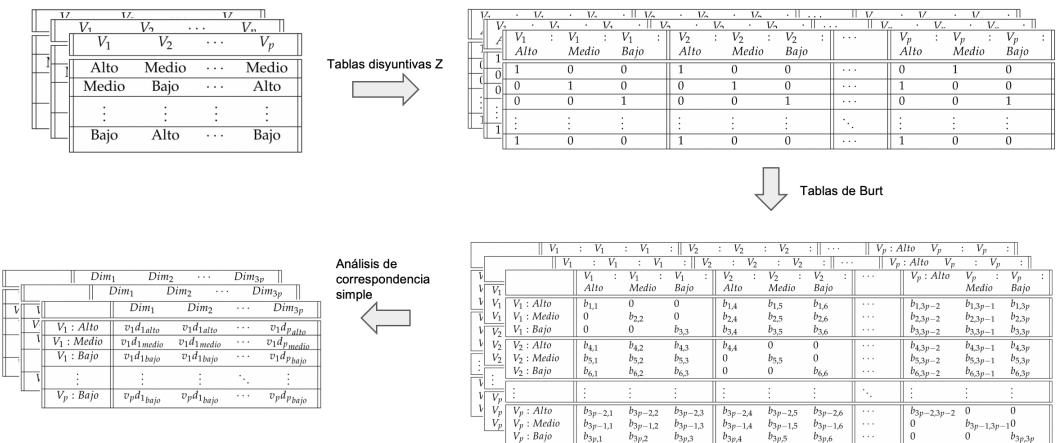
#### <sup>210</sup> 2.2.5. Aporte del Análisis Factorial Múltiple (MFA)

<sup>211</sup> Una vez que se tienen las coordenadas de las columnas, se procede a realizar la normalización,  
<sup>212</sup> característica del procedimiento MFA.

<sup>213</sup> Sea  $\lambda_1^k$  el primer valor propio obtenido de la descomposición singular de la  $k$ -ésima tabla  $C$ . Se  
<sup>214</sup> normaliza la tabla multiplicándola por  $1/\lambda_1^k$ . Con esto se obtiene la tabla  $C'$ , que corresponde a la tabla  
<sup>215</sup> de coordenadas normalizadas.

<sup>216</sup> Individualmente, para el caso de la matriz  $k$ , se tendría la siguiente expresión.

$$\mathbf{C}'_k = \frac{1}{\lambda_1^k} \mathbf{C}_k \quad (6)$$

Figure 2. Procedimiento del MCA para  $k$  tablas

217 Aglomerando las matrices normalizadas  $C'$  en una sola, se tiene la matriz  $\mathbb{C}'$ . Esta contiene todos  
 218 los elementos de las  $k$  tablas.

$$\mathbb{C}' = [\mathbf{C}_1' | \mathbf{C}_2' |, \dots, |\mathbf{C}_K']^T \quad (7)$$

219 La normalización que realiza el MFA se encarga de ponderar las  $k$  tablas, con el objetivo de evitar  
 220 alguna descompensación al momento de realizar el análisis conjunto de las tablas.

### 221 2.3. Gráfico de control $T2Qv$

#### 222 2.3.1. Obtención del gráfico de control

223 Para definir el gráfico de control  $T^2$  Hotelling se deben tomar las siguientes consideraciones:

- 224 • La tabla  $\mathbb{C}'$  (Ecuación 7) se denomina Consenso, sirve como referente para el escenario *bajo control*,  
 225 y de la cual se obtiene  $\mu_0$  y  $\mathbf{S}_0$ .
- 226
- 227 • Cada matriz  $\mathbf{C}_k'$  tiene el mismo número de filas ( $n$ ) y columnas ( $p$ ) (individuos y variables).
- 228 • El vector de medias  $\bar{\mu}_k$  está atado a la tabla  $\mathbf{C}_k'$ , es decir, el gráfico de control estará en función de  
 229 las diferencias entre las matrices  $\mathbf{C}_k'$  y la matriz consenso  $\mathbb{C}'$ .
- 230 • Las matrices  $\mathbf{C}_k'$  siguen una distribución normal multivariante con vector de medias  $\mu_k$  y matriz  
 231 de covarianzas  $\mathbf{S}_k$ .

232 Con esto se obtiene el estadístico  $T^2$ :

$$T^2 = n(\mu_k - \mu_0)' \Sigma_0^{-1} (\mu_k - \mu_0) \quad (8)$$

233 Se sabe que, bajo control, el  $T^2$  se distribuye como una Chi-cuadrado con  $p$  grados de libertad  $\chi^2_p$ .  
 234 En este caso se puede aplicar este principio, ya que se utiliza la matriz consenso ( $\mathbb{C}'$ ), que representa al  
 235 escenario bajo control.

236 Dado que este gráfico de control está basado en distancias de Mahalanobis ponderadas, sólo tiene  
 237 límite de control superior. Este viene dado por la ecuación 9

$$UCL = \chi_{\alpha', p}^2 \quad (9)$$

238 donde  $p$  es el número de dimensiones y  $\alpha'$  es la significancia predeterminada considerando  $p$ . Tal  
 239 que  $\alpha' = 1 - (1 - \alpha)^p$  y  $\alpha$ , que en este estudio se considera como el error tipo 1 inicial, generalmente es  
 240 igual a 0.0027.

241 Para definir  $\alpha'$  (Error tipo 1) se opta el enfoque múltiple, Montgomery [44] indica que con este  
 242 enfoque se consigue una variación de  $\alpha'$  en función de  $p$ , esto es conveniente en este caso ya que  
 243 permite considerar la dimensionalidad usada al realizar el análisis de correspondencias múltiples.

#### 244 2.3.2. Interpretación de puntos fuera de control

245 El gráfico multivariante  $T^2$  de Hotelling para variables cualitativas es capaz de señalar que el  
 246 proceso salió de control, pero no permite reconocer el momento ni las causas por las que ocurrió esto.  
 247 Es obvio que, más allá de reconocer el estado del proceso, interesa saber cuándo y por qué salió de  
 248 control. Es importante tener en cuenta que cada punto representado en el gráfico  $T^2$  de Hotelling  
 249 representa a una tabla (muestra), constituida por un grupo de individuos (observaciones) y  $p$  variables  
 250 que pueden tener muchas categorías, algunas de éstas pueden mostrar un comportamiento anómalo.  
 251 Por consiguiente, es necesario analizar con detenimiento qué está pasando con los datos de las tablas  
 252 reportadas.

253 Este análisis se realiza comparando la ubicación de los puntos que representan las categorías de  
 254 las variables en el MCA de la tabla consenso y la ubicación de los puntos en los gráficos MCA de  
 255 cada tabla reportada como fuera de control. Las categorías que están incidiendo en el estado fuera  
 256 de control son aquellas cuya ubicación en ambas tablas comparadas muestra diferencias importantes.  
 257 Para cuantificar la magnitud del comportamiento anómalo de estas categorías se calcula las distancias  
 258 Chi-cuadrado entre las masas de las columnas de la tabla reportada como fuera de control y las de  
 259 la tabla consenso, tomada como referente. Mientras mayor es el valor del estadístico, mayor es su  
 260 incidencia en el desplazamiento de la media del proceso que, finalmente, pueden llevarlo a un estado  
 261 fuera de control.

262 El aplicativo informático del gráfico T2Qv presenta esta información de dos maneras: la primera  
 263 es una tabla que registra las distancias  $\chi^2$  para cada categoría de las variables analizadas, además,  
 264 muestra el  $p$ -valor para cada observación, de esto depende el número de asteriscos que indican su  
 265 nivel de significancia estadística. Así, si el  $p$ -valor es inferior a 0.05, la observación se reporta como  
 266 significancia estadística y va asociada a un asterisco; si el  $p$ -valor es menor que 0.01, se entiende  
 267 que hay alta significación estadística y se registran dos asteriscos; si el  $p$ -valor es menor que 0.001,  
 268 la significancia estadística es muy alta y se reportan tres asteriscos; caso contrario, no se reporta  
 269 significancia y la observación no lleva asteriscos.

270 La segunda manera consiste en un gráfico de barras que incluye tres líneas horizontales  
 271 correspondientes a los límites asociados a los niveles de significancia estadística, la más baja representa  
 272 al  $p$ -valor inferior a 0.05 (un asterisco); la línea del medio, al  $p$ -valor inferior a 0.01 (dos asteriscos); y, la  
 273 línea más alta representa al  $p$ -valor inferior a 0.001 (tres asteriscos). Las barras que representan valores  
 274 sin significancia estadística no sobrepasan ninguna de las líneas y se pintan de color gris, mientras que,  
 275 las que sí denotan significancia estadística adquieren el color de la línea más alta que sobrepasan.

276 De esta manera, la metodología propuesta en esta investigación permite explicar cuándo y por  
 277 qué el proceso salió de control.

### 278 3. Complemento computacional

279 Para facilitar la difusión y aplicación del método propuesto, se ha desarrollado un paquete  
 280 reproducible en R. El paquete T2Qv utiliza la metodología expuesta en este artículo y la lleva a un  
 281 entorno práctico, permite visualizar los resultados de forma plana o interactiva, además, presenta un  
 282 panel Shiny que contiene todas las funciones individuales en un mismo espacio.

#### 283 3.1. Disponibilidad

284 El paquete está disponible en GitHub, la descarga se la puede realizar de la siguiente forma:

```
285 install.packages("devtools")
286 devtools::install_github("JavierRojasC/T2Qv")
```

<sup>287</sup> 3.2. El paquete: T2Qv

```

Package: T2Qv
Type: Package
Title: Control Qualitative Variables
Version: 0.1.0
Authors@R: c(person("Wilson", "Rojas-Preciado", role = c("aut", "cre"),
  email = "wrojas@utmachala.edu.ec"),
person("Mauricio", "Rojas-Campuzano", role = c("aut", "ctb"),
  email="maurijoja@espol.edu.ec"),
person("Purificación", "Galindo-Villardón", role = c("aut", "ctb"),
  email = "oruij@espol.edu.ec"),
person("Omar", "Ruiz-Barzola", role = c("aut", "ctb"),
  email = "oruij@espol.edu.ec"))
Maintainer: Wilson Rojas-Preciado <wrojas@utmachala.edu.ec>
Description: Covers k-table control analysis using multivariate control charts for qualitative variables using
fundamentals of multiple correspondence analysis and multiple factor analysis. The graphs can be shown in a
flat or interactive way, in the same way all the outputs can be shown in an interactive shiny panel.
License: MIT + file LICENSE
Encoding: UTF-8
LazyData: true
RoxygenNote: 7.1.1
Depends: R (>= 2.10)
Imports: shiny, shinydashboardPlus, shinydashboard, shinycssloaders,
  dplyr, ca, highcharter, stringr, tables, htmltools (>= 0.5.1.1)
Suggests: testthat (>= 3.0.0)
Config/testthat/editon: 3
Author: Wilson Rojas-Preciado [aut, cre],
  Mauricio Rojas-Campuzano [aut, ctb],
  Purificación Galindo-Villardón [aut, ctb],
  Omar Ruiz-Barzola [aut, ctb]
Built: R 4.0.2 ; 2021-10-14 23:56:56 UTC; unix
  
```

**Figure 3.** Documentación del paquete T2Qv

<sup>288</sup> Las funciones que contiene el paquete y su descripción se enuncian en la tabla 9.

| Función        | Descripción  |
|----------------|--|
| T2 qualitative | Multivariate control chart T2 Hotelling applicable for qualitative variables.  |
| MCAconsensus   | Multiple correspondence analysis applied to a consensus table.   |
| MCApoint       | Multiple correspondence analysis applied to a specific table.  |
| ChiSq variable | Contains Chi square distance between the column masses of the table specified in PointTable and the consensus table. It allows to identify which mode is responsible for the anomaly in the table in which it is located.          |
| Full Panel     | A shiny panel complete with the multivariate control chart for qualitative variables, the two MCA charts and the modality distance table. Within the dashboard, arguments such as type I error and dimensionality can be modified. |

**Table 9.** Funciones del paquete T2Qv

<sup>289</sup> 4. Resultados

<sup>290</sup> Con la intención de probar la metodología propuesta en el gráfico de control  $T^2$  de Hotelling  
<sup>291</sup> para variables cualitativas, se hizo un análisis con datos simulados y otro con datos reales aplicados al  
<sup>292</sup> contexto de la educación superior. Los resultados se obtienen de la aplicación del paquete T2Qv.

<sup>293</sup> 4.1. Resultados con datos simulados

<sup>294</sup> Para este estudio se generó una base de datos simulados, a la que se denominó  
<sup>295</sup> *Datak10Contaminated*. Consta de 10 tablas, cada una de ellas está constituida por 100 filas  
<sup>296</sup> (observaciones) y 11 columnas, de las cuales, las 10 primeras corresponden a las variables analizadas  
<sup>297</sup> (V1, V2, ...; V10) mientras que, la columna 11, denominada *GroupLetter*, contiene el factor de  
<sup>298</sup> clasificación de los grupos. Para su identificación, las tablas han sido denominadas con las letras del  
<sup>299</sup> alfabeto, desde la *a* hasta la *j*. La tabla *j* tiene una distribución distinta de la que tienen las otras nueve.  
<sup>300</sup> Los datos se expresan en tres niveles: alto, medio y bajo. La tabla ?? presenta las 10 primeras filas de la  
<sup>301</sup> base de datos *Datak10Contaminated*.

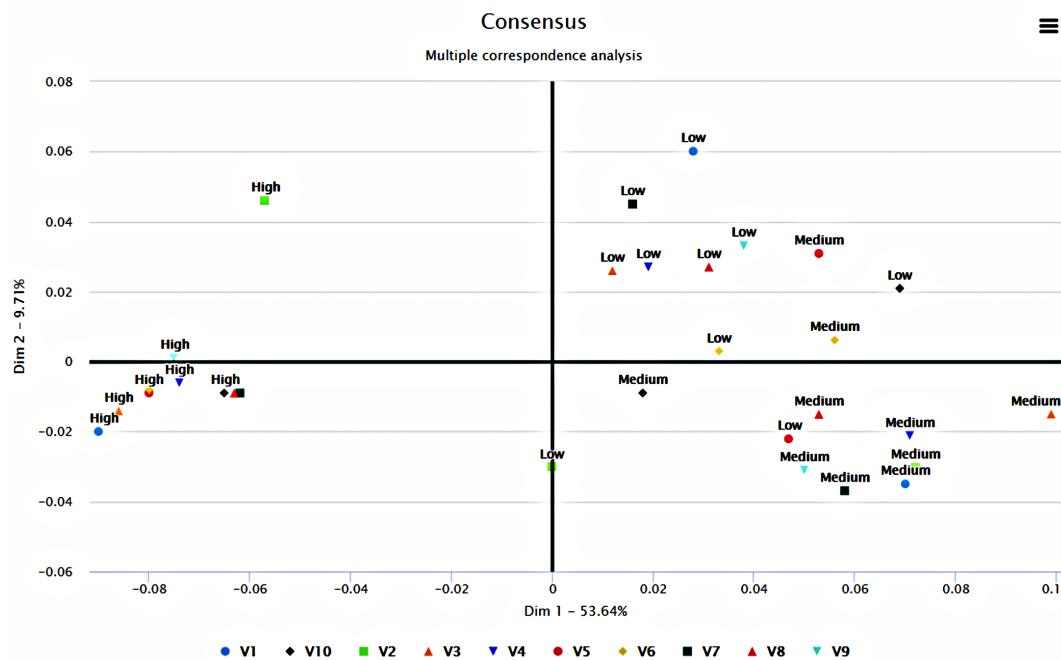
| V1     | V2     | V3     | V4     | V5     | V6     | V7     | V8     | V9     | V10    | GroupLetter |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|-------------|
| Low    | Medium | Medium | High   | High   | High   | Low    | Medium | Medium | Medium | a           |
| Low    | Low    | High   | Low    | Medium | High   | High   | High   | Low    | High   | a           |
| High   | Medium | High   | Low    | High   | Medium | Medium | High   | Medium | Low    | a           |
| Medium | Medium | Low    | High   | Low    | Medium | High   | Low    | Low    | High   | a           |
| Low    | Low    | Low    | High   | Low    | High   | High   | High   | Medium | Medium | a           |
| High   | High   | Medium | Low    | Low    | Low    | Medium | Medium | High   | Low    | a           |
| High   | High   | Low    | Low    | Low    | Medium | High   | Medium | Medium | High   | a           |
| Medium | Medium | High   | Medium | Medium | High   | Medium | High   | High   | High   | a           |
| Low    | Low    | Low    | Medium | High   | Medium | Low    | Medium | Low    | Low    | a           |
| Medium | Medium | Medium | High   | Low    | Medium | High   | Low    | High   | Medium | a           |

**Table 10.** Sección de la base de datos Datak10Contaminated.

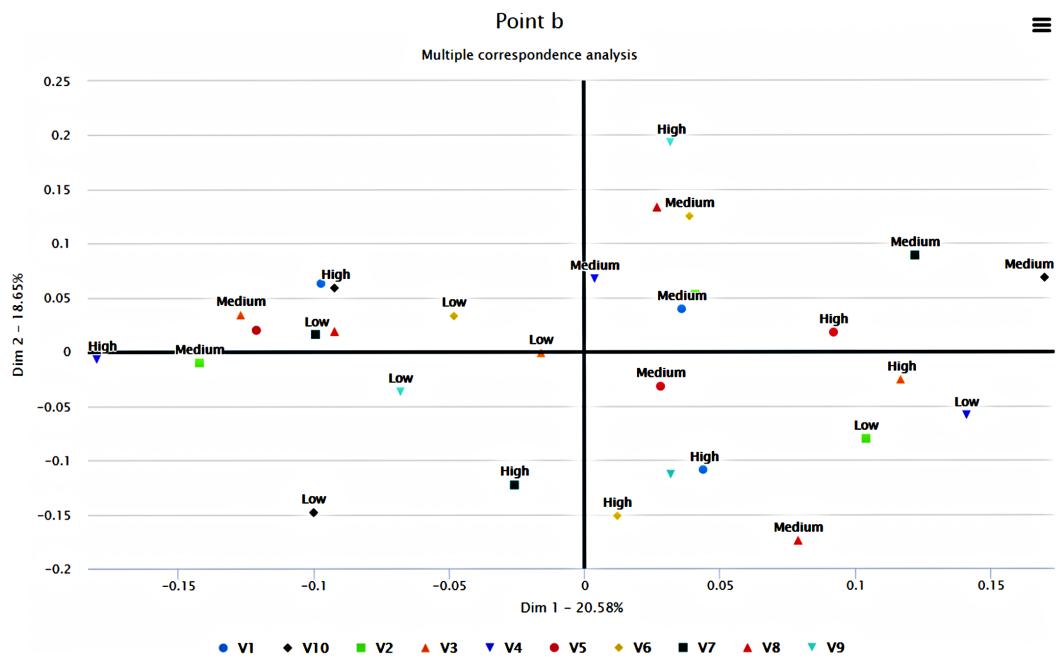
302 Para facilitar análisis se creó un paquete al que se denominó *T2Qv*, herramienta diseñada en el  
 303 software estadístico R y R Studio. *T2Qv* realiza el análisis de control de  $k$  tablas utilizando gráficos  
 304 de control multivariantes para variables cualitativas, utilizando los fundamentos del análisis de  
 305 correspondencias múltiples y el análisis de factores múltiples. Los gráficos se pueden mostrar de forma  
 306 plana o interactiva, de la misma manera todas las salidas se pueden mostrar en un panel interactivo de  
 307 Shiny.

308 El primer resultado es el gráfico del Análisis de Correspondencias Múltiples (MCA) aplicado a la tabla  
 309 consenso (Figura 4). Esta tabla ha sido tomada como referente, como escenario en control para el  
 310 análisis posterior de las tablas que sean reportadas como puntos fuera de control en el gráfico  $T^2$  de  
 311 Hotelling.

312 El MCA reporta una inercia total del 63.3%, la dimensión 1 representa al 53.6% de la información,  
 313 mientras que la dimensión 2, al 9.7%. Los puntos del gráfico representan a las observaciones de cada  
 314 una de las 10 variables en sus tres niveles: alto, medio y bajo. En esta figura, todas las observaciones  
 315 que corresponden al nivel alto se ubican a la izquierda en el eje de las X; de las 10 observaciones  
 316 correspondientes al nivel medio, 8 se situaron en el cuarto cuadrante y las dos restantes en el cuadrante  
 317 1, es decir, todas las observaciones de este nivel estuvieron a la derecha en el eje de las X. Finalmente,  
 318 de los 10 puntos que representan al nivel bajo, 8 están ubicados en el cuadrante 1.

**Figure 4.** Análisis de correspondencias múltiples aplicado a la tabla consenso.

319      Otro resultado es el Análisis de Correspondencias Múltiples aplicado a una tabla específica. En  
 320 este punto, uno de los argumentos que se debe tener en cuenta es la selección de la tabla de la que se  
 321 realizará el análisis.



**Figure 5.** Análisis de correspondencias múltiples aplicado a la tabla b.

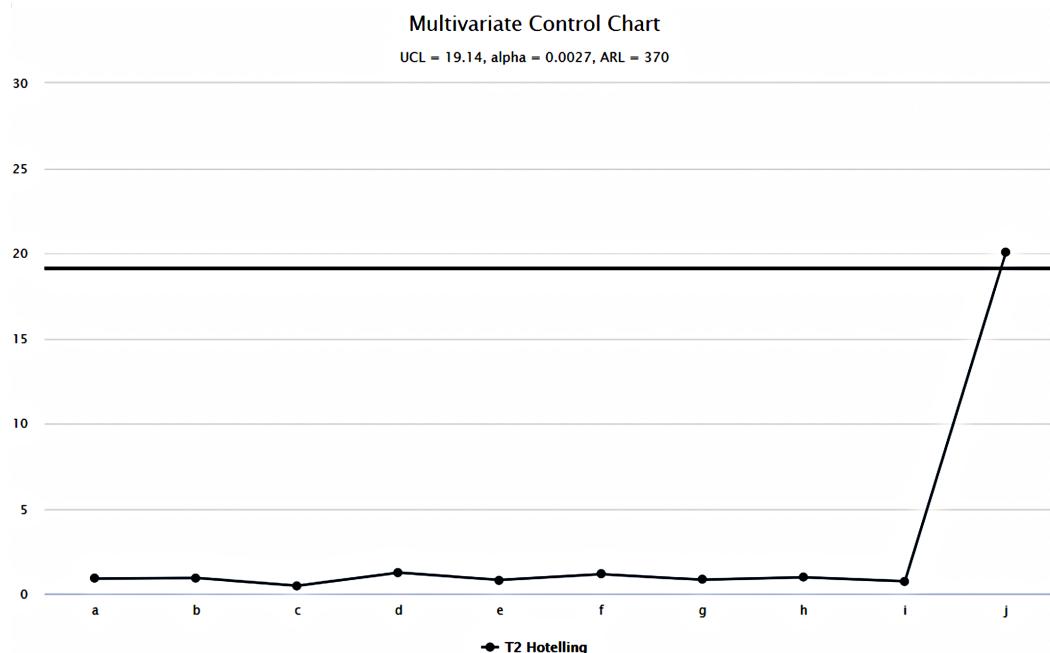
322      La figura 5 representa el gráfico del MCA de la tabla b. Este gráfico, en sus dos dimensiones,  
 323 representa al 39.3% de la información. Es notorio que las observaciones en sus niveles alto, medio  
 324 y bajo están distribuidas de forma aleatoria en todos los cuadrantes del del gráfico, no se puede  
 325 precisar un patrón específico de agrupación. Esto mismo se puede decir de los puntos representados en  
 326 cualquiera de las otras tablas, exceptuando la tabla j, que fue diseñada con una distribución diferente.  
 327 No obstante, el uso del MCA de las figuras 5 y 4 todavía no permite detectar si el proceso está o no  
 328 en control. La identificación de puntos fuera de control se puede realizar mediante la representación  
 329 gráfica del estadístico  $T^2$  de Hotelling, como se observa en la figura 6.

330      La figura 6 presenta un gráfico de control elaborado con el estadístico  $T^2$  de Hotelling, aplicado  
 331 a la detección de anomalías en cualquiera de las k tablas analizadas. Cada una de las tablas está  
 332 representada por los puntos en el gráfico. Se observa una línea horizontal que representa al límite de  
 333 control superior (UCL). El límite de control inferior (LCL) es igual a cero.

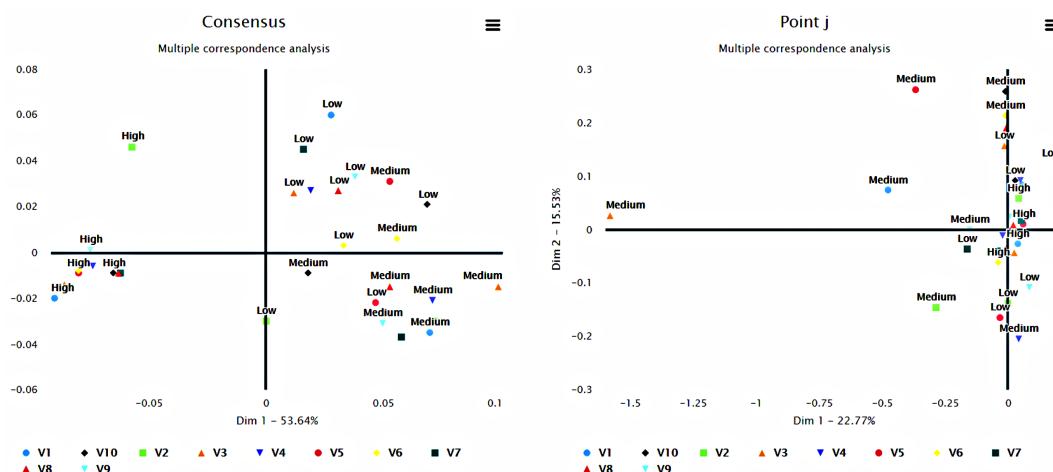
334      Dado que el análisis de sensibilidad determinó que este gráfico de control tiene un mejor rendimiento  
 335 cuando trabaja con un número alto de dimensiones, se ha recomendado que este sea  $p-1$ , donde p es el  
 336 número de dimensiones inicial, que es equivalente a la cantidad de variables de la base de datos, sin  
 337 contar a la variable GroupLetter que sólo sirve como factor de clasificación de las tablas.

338      Se observa que el punto que representa a la tabla j se ubica por encima del límite de control superior, lo  
 339 que quiere decir que se lo ha identificado como un valor fuera de control. Por consiguiente, es necesario  
 340 analizar con detenimiento qué está pasando con los datos de la tabla reportada, comparándolos con  
 341 los de la tabla consenso, a fin de identificar las causas de la variación y tomar las acciones pertinentes.  
 342 Para hacer un análisis del punto fuera de control se realiza un gráfico del MCA de la tabla j y se lo  
 343 compara con el gráfico similar de la tabla consenso, como se presenta en la figura 7.

344      La figura 7 presenta la distribución de las observaciones de las tablas consenso y j mediante  
 345 gráficos del MCA. El gráfico de la tabla consenso, que sirve de referente en control, ya se analizó en  
 346 la figura 4; el de la tabla j muestra una tendencia de los puntos que con valores medios a ubicarse al  
 347 lado izquierdo, bastante alejados de los demás que confluyen hacia el centro del eje de las X. Especial



**Figure 6.** Gráfico de control multivariante T2 Hotelling aplicable a variables cualitativas.



**Figure 7.** Comparación de los gráficos del MCA aplicado a la tabla consenso y la tabla j.

348 atención merece la variable 3, que registra una observación para el nivel medio con el valor más alejado  
349 del grupo.

350 Al comparar los gráficos es obvio que la distribución de los datos en la tabla *j* es diferente de las  
351 distribuciones de las demás tablas, y en especial, es diferente de la distribución de la tabla consenso, lo  
352 que explica por qué el punto *j* ha sido identificado como fuera de control.

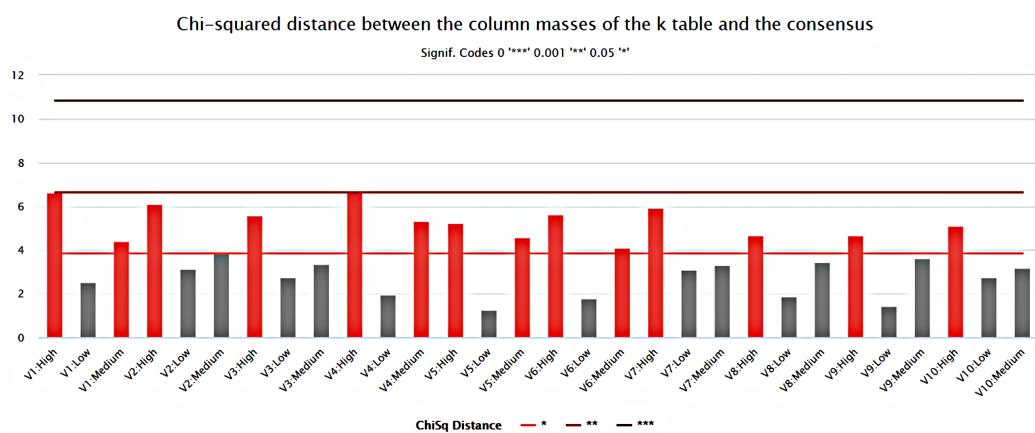
353 La tabla 11 contiene los datos de cada una de las variables de la tabla *j* con sus tres niveles (alto,  
354 medio y bajo). La columna 3 muestra el p-valor para cada observación, de esto depende el número de  
355 asteriscos de la columna 4 que indica el nivel de significancia estadística.

356 De las 30 observaciones que tiene la tabla *j*, se presentan 14 casos de p-valores menores que 0.05,  
357 es decir, reportan significancia estadística (un asterisco), de los cuales, 10 se atribuyen a las categorías  
358 altas de las variables cualitativas y cuatro a los niveles medios. El comportamiento de estas variables en  
359 la tabla *j*, que obedece a una distribución diferente a la de las demás tablas, provoca el desplazamiento

| Variable   | Chi-Squared | p-value | Signif |
|------------|-------------|---------|--------|
| V1:High    | 6.62953     | 0.01003 | *      |
| V1:Low     | 2.51447     | 0.11281 |        |
| V1:Medium  | 4.40573     | 0.03582 | *      |
| V2:High    | 6.10216     | 0.01350 | *      |
| V2:Low     | 3.15682     | 0.07561 |        |
| V2:Medium  | 3.81899     | 0.05067 |        |
| V3:High    | 5.58957     | 0.01807 | *      |
| V3:Low     | 2.73051     | 0.09845 |        |
| V3:Medium  | 3.37596     | 0.06615 |        |
| V4:High    | 6.61362     | 0.01012 | *      |
| V4:Low     | 1.95916     | 0.16160 |        |
| V4:Medium  | 5.33225     | 0.02093 | *      |
| V5:High    | 5.23785     | 0.02210 | *      |
| V5:Low     | 1.24566     | 0.26438 |        |
| V5:Medium  | 4.56461     | 0.03264 | *      |
| V6:High    | 5.64217     | 0.01753 | *      |
| V6:Low     | 1.81050     | 0.17845 |        |
| V6:Medium  | 4.11597     | 0.04248 | *      |
| V7:High    | 5.94284     | 0.01478 | *      |
| V7:Low     | 3.10801     | 0.07791 |        |
| V7:Medium  | 3.32453     | 0.06825 |        |
| V8:High    | 4.65021     | 0.03105 | *      |
| V8:Low     | 1.88624     | 0.16963 |        |
| V8:Medium  | 3.45642     | 0.06301 |        |
| V9:High    | 4.67660     | 0.03058 | *      |
| V9:Low     | 1.46059     | 0.22684 |        |
| V9:Medium  | 3.60793     | 0.05750 |        |
| V10:High   | 5.10688     | 0.02383 | *      |
| V10:Low    | 2.76592     | 0.09629 |        |
| V10:Medium | 3.19604     | 0.07382 |        |

**Table 11.** Distancia  $\chi^2$  entre las masas de la tabla consenso y las  $k$ , Datak10Contaminated

360 de la media del proceso que, al final, lo lleva a un estado fuera de control. Otra manera de visualizar  
 361 esta información es a través de un gráfico de barras (figura 8).



**Figure 8.** Distancia  $\chi^2$  entre las masas de la tabla consenso y las  $k$  tablas, Datak10Contaminated.

362 El gráfico de barras de la figura 8, expresa también la distancia  $\chi^2$  entre las masas de la tabla  
 363 consenso y las  $k$  tablas de la base de datos Datak10Contaminated. Es otra manera de representar la  
 364 información que ya se analizó en la tabla 11.

365 4.2. Resultados con datos aplicados al contexto de la educación superior

366 En este ejemplo se utiliza una base de datos denominada *Estudiantes\_2019\_2020*, tomada de  
367 reportes que la Universidad Técnica de Machala (UTMACH) cargó en la plataforma del Sistema Integral  
368 de Información de la Educación Superior (SIIES), correspondiente a cuatro periodos académicos  
369 consecutivos. La base de datos *Estudiantes\_2019\_2020* contiene 43191 observaciones y 17 variables  
370 cualitativas referidas los estudiantes de las 30 carreras vigentes en sus 5 facultades.

371 Las variables registradas en la base de datos, con sus respectivas categorías son las siguientes:

- 372 • Periodo académico, esta es la variable que sirve como clasificador, hace referencia a 4 periodos de  
373 estudio (semestres): 2019-1, 2019-2, 2020-1 y 2020-2.
- 374 • Facultad, que tiene 5 categorías: Facultad de Ciencias Agropecuarias (FCA), Facultad de Ciencias  
375 Empresariales (FCE), Facultad de Ciencias Químicas y de la Salud (FCQS), Facultad de Ciencias  
376 Sociales (FCS) y Facultad de Ingeniería Civil (FIC).
- 377 • Carrera, variable que contiene el nombre de las 30 carreras vigentes en la UTMACH, cada  
378 una de ellas es una categoría y se asocia a alguna de las 5 facultades. En la FCA: Acuicultura,  
379 Economía Agropecuaria, Agronomía y Medicina Veterinaria; en la FCE: Administración de  
380 Empresas, Turismo, Mercadotecnia, Contabilidad y auditoría, Comercio internacional y Economía;  
381 en la FCQS: Medicina, Enfermería, Bioquímica y Farmacia, Alimentos, Ing. Química; en la  
382 FCS: Artes plásticas, Pedagogía de la Actividad Física y Deporte, Pedagogía de las Ciencias  
383 Experimentales, Educación Básica, Educación inicial, Pedagogía de los Idiomas Nacionales y  
384 Extranjeros, Psicología Clínica, Psicopedagogía, Comunicación, Derecho, Gestión Ambiental,  
385 Sociología, Trabajo Social; y en la FIC: Ingeniería Civil y Tecnología de la Información.
- 386 • Sexo, con sus dos clases: hombre y mujer.
- 387 • Grupo edad, que clasifica a los estudiantes en 5 grupos según su edad en años: Menores que 18,  
388 de 18 a 30, de 31 a 45, de 46 a 60 y Mayores a 60.
- 389 • Discapacidad, cuyas clases son: Intelectual, Auditiva, Física Motora, Visual, Lenguaje y Ninguna.
- 390 • Etnia, con sus tipos: Mestizo, Montubio, Negro, Blanco, Indígena, Mulato, Afroecuatoriano, Otro,  
391 No registra.
- 392 • Zona residencial, Urbana y Rural.
- 393 • Nivel de formación del padre: Centro de alfabetización, Educación Básica incompleta,  
394 Educación Básica, Bachillerato, Superior tecnológica incompleta, Superior tecnológica, Superior  
395 universitaria, Superior universitaria incompleta, Diplomado, Especialidad, Postgrado Maestría o  
396 Especialización en áreas de Salud, Postgrado Ph.D., Ninguno y No sabe, no registra.
- 397 • Nivel de formación de la madre: Centro de alfabetización, Educación Básica incompleta,  
398 Educación Básica, Bachillerato, Superior tecnológica incompleta, Superior tecnológica, Superior  
399 universitaria, Superior universitaria incompleta, Diplomado, Especialidad, Postgrado Maestría o  
400 Especialización en áreas de Salud, Postgrado Ph.D., Ninguno y No sabe, no registra.
- 401 • Número de miembros del hogar, con sus tres clases: Hasta 3, 4 y 5 o más.
- 402 • Tipo colegio: Fiscal, Particular, Fiscomisional, Extranjero, Municipal y No registra.
- 403 • Ingreso total hogar: Rango 1, Rango 2, Rango 3, Rango 4, Rango 5, Rango 6, Rango 7, Rango 8,  
404 Rango 9 y Rango 10.
- 405 • Origen de recursos estudios: Padres tutores, Hermanos y familiares, Pareja sentimental, Recursos  
406 propios, Beca estudio, Crédito educativo y No registra.
- 407 • Segunda matrícula: Sí y No.
- 408 • Tercera matrícula: Sí y No.
- 409 • Terminó periodo: Sí y No.

410 4.2.1. Análisis de Correspondencias Múltiples de la tabla Consenso

411 El gráfico del Análisis de Correspondencias Múltiples que se realiza a la tabla consenso (Figura 4)  
412 es el escenario bajo control que se utilizará para el análisis de las tablas que luego se registren como  
413 puntos fuera de control en el gráfico  $T^2$  de Hotelling. El MCA reporta una inercia total del 30.31%. Los

414 puntos del gráfico representan a las observaciones de cada una de las 16 variables en sus distintos  
 415 niveles. La variable Periodo académico sirve como elemento clasificador, por eso sus observaciones no  
 416 aparecen aquí.

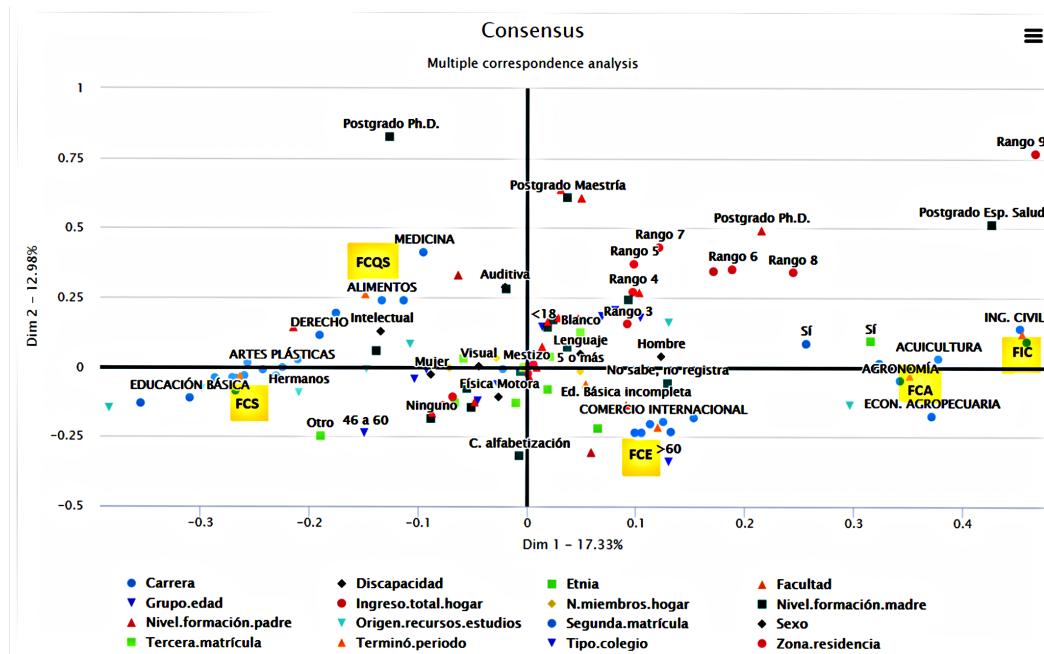


Figure 9. Gráfico de MCA de la tabla consenso, Estudiantes 2019-2020.

417 Se observa cómo las carreras se agrupan alrededor de sus respectivas facultades; la FIC y la FCA  
 418 se muestran similares entre sí y ubicadas al lado derecho, en el plano que corresponde a la dimensión  
 419 1, mientras que, al otro extremo está la FCS. Por otra parte, las similitudes y diferencias entre las otras  
 420 dos facultades giran en torno a los ejes de las dos dimensiones, la FCE ubicada en el cuadrante 4 y la  
 421 FCQS, en el 2.

422 La variable Sexo es una de las que más incide en la ubicación de los puntos alrededor de la dimensión  
 423 1. El número de estudiantes varones es mayor que el de las mujeres en las carreras de la FCA y FIC, por  
 424 otra parte, el número de mujeres es mayor que el de hombres en las carreras de la FCS y FCQS; en la  
 425 FCE parece no haber marcada diferencia en la proporción de hombres y mujeres. Las variables Segunda  
 426 matrícula y Tercera matrícula dan cuenta de que es muy frecuente que los estudiantes aprueben sus  
 427 asignaturas en su primera matrícula, sin repetir; se observa también que la segunda y tercera matrícula  
 428 ocurren con mayor frecuencia en la FCA y la FIC, especialmente en ésta, lo que podría estar asociado al  
 429 grado de dificultad propio de las asignaturas que allí se estudian, a procesos con mayor rigor académico  
 430 y hasta a insuficiencias en los procesos de enseñanza – aprendizaje. Al otro extremo está la FCS, en la  
 431 que no es usual que ocurran segundas o terceras matrículas.

432 La variable Ingreso total hogar se desplaza desde el nivel más bajo (Rango 1), que se encuentra cercano  
 433 a las carreras de la FCS, FCQS, hasta los más altos, que corresponden a las carreras de la FCA y FIC. Los  
 434 valores medios - altos (Rangos 5, 7) están cercanos a la carrera de Medicina en la FCQS. Las carreras  
 435 del área social son preferidas por estudiantes que provienen de familias con bajos ingresos, lo cual es  
 436 congruente con la observación de que las becas de estudio, de la variable Origen recursos estudio, se  
 437 han direccionado de manera preferente a estudiantes de la FCS.

438 Por otra parte, se observa que la mayoría de los estudiantes de la universidad reside en zonas urbanas,  
 439 pero la categoría Zona rural se acerca más a las carreras de la FCS. Además, los estudiantes de la  
 440 FCS y FCE provienen, mayoritariamente, de colegios fiscales y municipales; los niveles de formación  
 441 académica de padres y madres de los estudiantes son más bajos en estos grupos, donde es usual  
 442 encontrar casos de educación básica incompleta, formación en centros de alfabetización y ninguna

443 formación. Los estudiantes que provienen de colegios particulares y fiscomisionales están con mayor  
 444 frecuencia en las carreras de la FCQS, FCA y FIC; asimismo, los niveles más altos de formación de  
 445 los padres y madres, como Postgrado Ph. D, Maestrías y Especializaciones médicas se asocian a  
 446 carreras como Medicina e Ingeniería Civil. El análisis de las variables que se manifiestan con mayor  
 447 presencia en la dimensión 2 del MCA indica que los grupos de estudiantes más jóvenes prefieren  
 448 carreras relacionadas con las ciencias médicas y de la salud, ciencias naturales y exactas, ingenierías y  
 449 tecnologías y ciencias agrícolas; mientras que, los grupos de mayor edad se asocian a carreras que se  
 450 ubican en el área de las ciencias sociales y las humanidades. La FCA y FIC reportan menor frecuencia  
 451 de casos de estudiantes con discapacidades que las demás facultades.

452 4.2.2. Gráfico de control multivariante T2 Hotelling

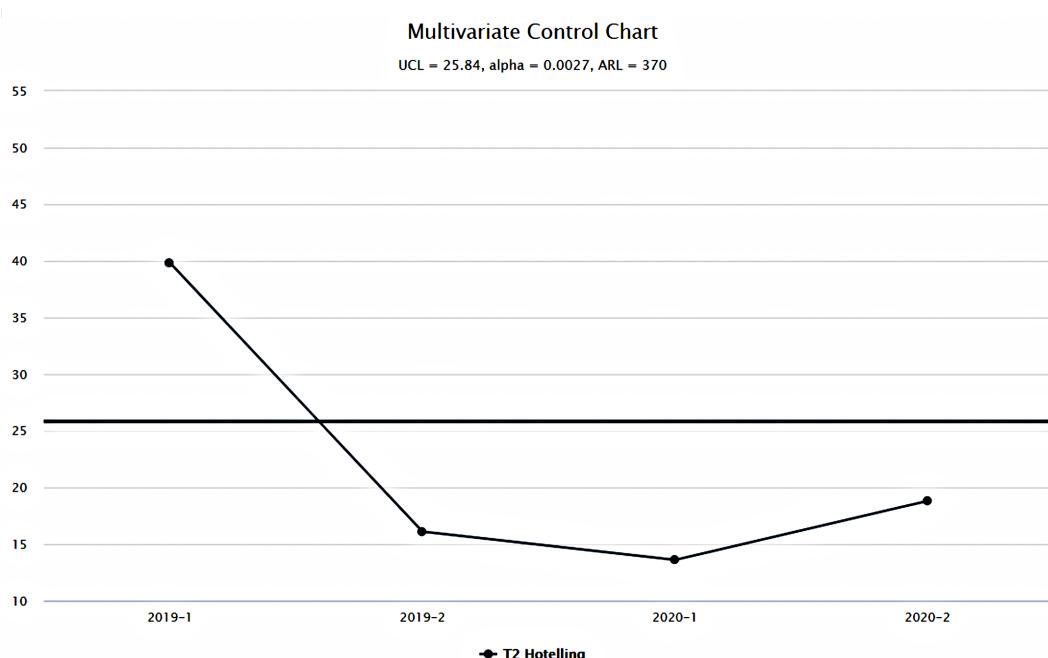
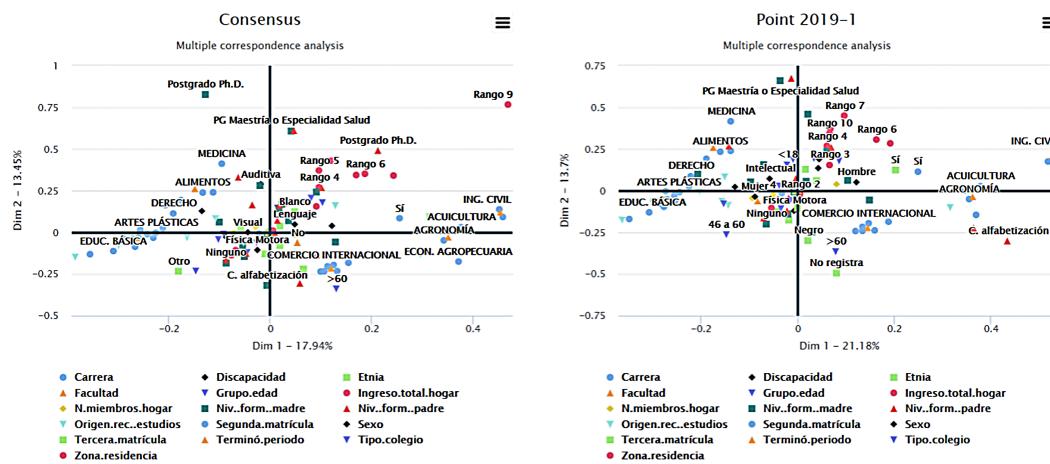


Figure 10. Gráfico de control T2 Hotelling aplicado a las tablas de los períodos académicos analizados.

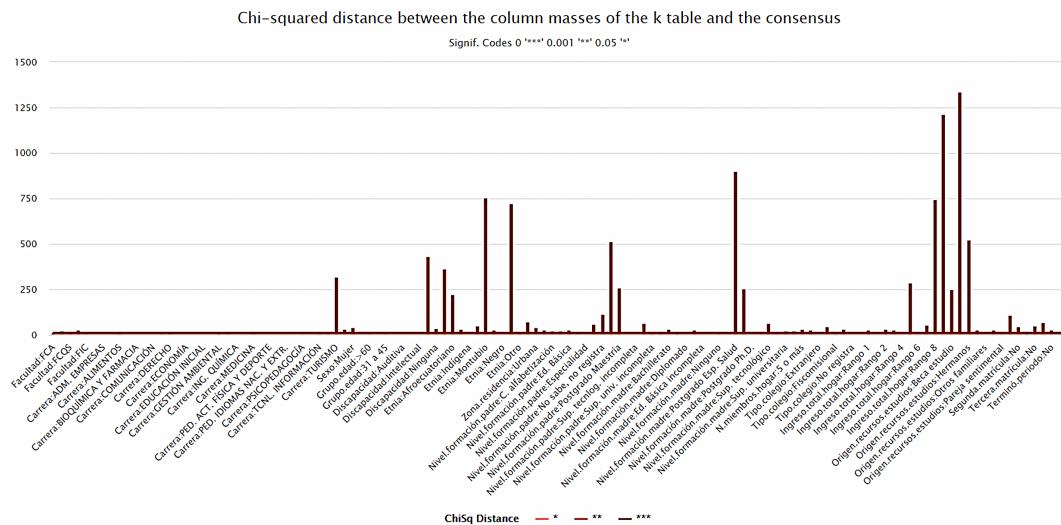
453 La figura 10 muestra el gráfico de control  $T^2$  de Hotelling para la representación de las  $k = 4$   
 454 tablas analizadas, éstas se representan por los puntos del gráfico y corresponden a los cuatro períodos  
 455 académicos considerados en este estudio. El punto que representa al período académico 2019-1 ha sido  
 456 reportado como un valor fuera de control, en consecuencia, será necesario un análisis de sus datos  
 457 comparados con los de la tabla consenso (Figura 9) para identificar las causas de la variación y, si fuera  
 458 el caso, tomar las acciones pertinentes. Para ello se realiza un MCA a la tabla 2019-1.

459 La figura 12 contiene los gráficos del MCA de la tabla consenso y la tabla 2019-1. Los puntos allí  
 460 registrados corresponden a las 16 variables cualitativas con sus respectivas categorías. Se ve que hay  
 461 puntos que conservan su ubicación o que han variado muy poco en ambas tablas, como las facultades,  
 462 carreras, la variable Sexo y la Zona residencia. Además, se observa categorías de variables que han  
 463 cambiado su ubicación de manera sensible y que pueden estar ocasionando el estado fuera de control.  
 464 La identificación de estas tablas se facilita cuando se analiza la figura 12.

465 La figura 12, permite apreciar, en un gráfico de barras, la distancia Chi cuadrado entre las  
 466 categorías de la tabla consenso y de la tabla 2019-1, reportada como fuera de control. Mientras más  
 467 altas son las barras, mayor es esta distancia. Las barras que sobresalen representan a las variables que,  
 468 en la comparación, tienen una distribución muy distinta, de manera que han alcanzado significancia  
 469 estadística muy alta y p-valores inferiores a 0.001 (tres asteriscos), por eso han adoptado el color  
 470 correspondiente a ese nivel en el gráfico. Estas variables son las que con mayor fuerza están provocando



**Figure 11.** Comparación de los gráficos del MCA aplicado a la tabla consenso y la tabla 2019-2.

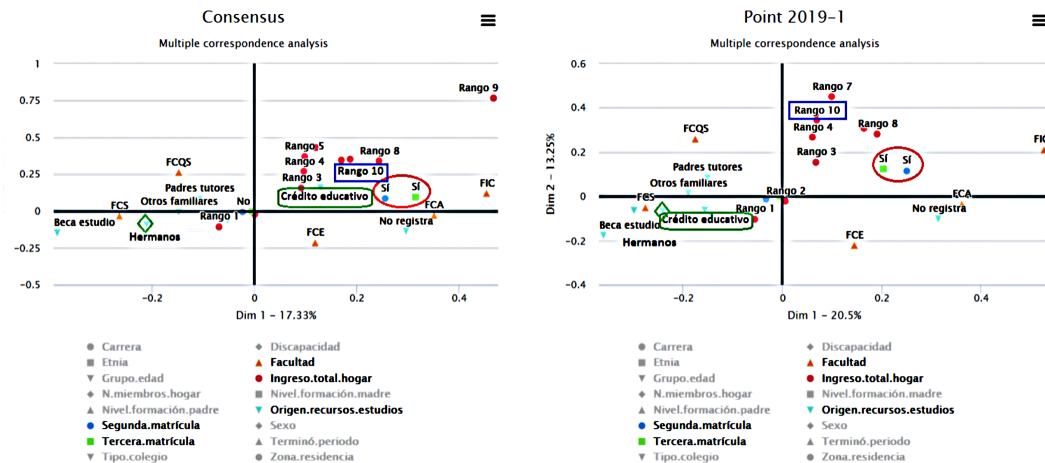


**Figure 12.** Distancia  $\chi^2$  entre las masas de la tabla consenso y las  $k$  tablas, Estudiantes 2019 2020.

el desplazamiento de la media del proceso y llevando al punto a un estado fuera de control. En consecuencia, es en ellas que se debe profundizar el análisis comparativo mediante el MCA.

En la figura 13, la barra más alta representa a la categoría Rango 9 de la variable Ingreso total hogar, que se ubica en la esquina superior del cuadrante 1 en la tabla consenso, pero, ya no aparece en la tabla 2019-1. Las categorías Rango 8 y Rango 10 tuvieron un desplazamiento hacia la izquierda. Durante el periodo de estudio, los estudiantes que provienen de familias con ingresos más altos han ido migrando desde Ingeniería Civil, Tecnologías de la Información, Acuicultura, Medicina Veterinaria y Agronomía, hacia carreras como Medicina e Ingeniería Química; los de ingresos medios se van alejando de las carreras como Administración de Empresas, Economía, Contabilidad y Auditoría, mientras que los estudiantes de bajos recursos no han modificado su preferencia: Carreras de Educación, Sociología y Trabajo Social.

Otra variable que demuestra alta incidencia en el desplazamiento de la media del proceso es Origen recursos estudios. Se observa que la categoría Crédito educativo demuestra un desplazamiento considerable, en la tabla consenso se ubica en el primer cuadrante, mientras que, en la tabla 2021-1 aparece en el tercero (figura 13). Esto implica que el crédito educativo que se ofrece a los estudiantes ha cambiado de dirección, desde áreas relacionadas con las ciencias sociales y humanísticas hasta áreas administrativas, ingenierías, tecnologías y ciencias agrícolas. Por otra parte, la categoría Hermanos



**Figure 13.** Comparación de los gráficos del MCA con énfasis en las variables de mayor significancia estadística.

488 y familiares se mantiene cerca de las carreras de la FCS, lo que significa que los estudiantes de bajos  
 489 recursos que estudian carreras de áreas sociales y humanísticas obtienen ayuda económica de sus  
 490 familiares.

491 El nivel Postgrado Ph.D. de las variables Nivel de formación del padre y Nivel de formación de la  
 492 madre manifiesta diferencias altamente significativas en los dos gráficos de MCA, pues sólo aparece en  
 493 la tabla consenso, no en la 2019-1. Esto se entiende porque a principios de 2019 todavía no había padres  
 494 de familia de la UTMACH que ostentaran ese grado académico, y en la comunidad general eran pocos.  
 495 No obstante, con el transcurso del tiempo varios de ellos, que estaban en proceso de formación de  
 496 doctorado, han logrado titularse, además, otros padres de familia que ya tenían tal grado académico  
 497 han matriculado a sus hijos en esta universidad.

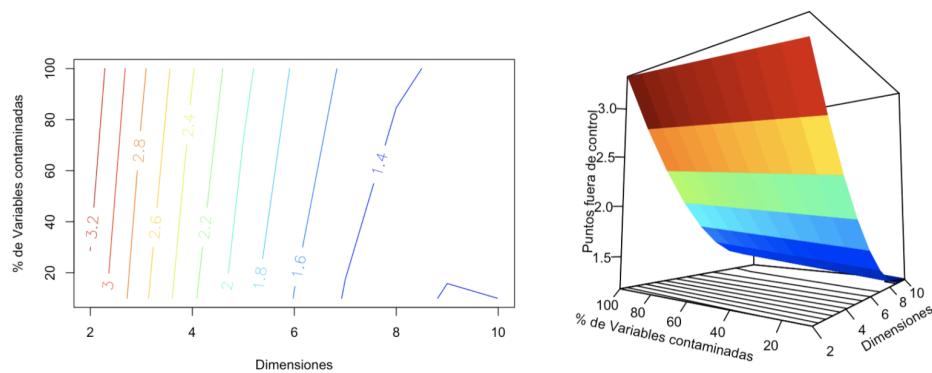
498 Al hacer un análisis del comportamiento de la variable Etnia en los años 2019 - 2020, llama la atención  
 499 que estudiantes que se autodefinieron como Negros, Afroecuatorianos y Mulatos, se alejan de carreras  
 500 de áreas sociales y humanísticas para acercarse a otras del área administrativa, como Administración  
 501 de Empresas, Contabilidad y Auditoría, Comercio internacional, Turismo, Mercadotecnia y Economía.  
 502 Mientras que, estudiantes que se autodenominaron Indígenas, migraron desde éstas hacia otras  
 503 carreras en áreas sociales y humanísticas.

504 Como cambios relevantes en torno a la variable Discapacidad se tiene que, en los períodos académicos  
 505 2019-1 hasta 2020-2, la discapacidad Intelectual se acerca a la carrera de Derecho; la discapacidad  
 506 Auditiva, a la Ingeniería química, Alimentos y Medicina. Mientras tanto, disminuye la frecuencia de  
 507 estudiantes con discapacidad Visual en Gestión Ambiental, Artes plásticas y Pedagogía de la Actividad  
 508 Física y Deporte.

## 509 5. Análisis de sensibilidad

510 Como se ha mencionado, en el gráfico T2Qv un punto fuera de control se interpreta como una  
 511 tabla ( $k_i$ ) que incluye una cantidad o una proporción de variables contaminadas, de tal manera que  
 512 la diferencia de los valores de masas de columna, entre de la matriz  $k_i$  y la matriz consenso, sean  
 513 significativos según el valor  $p$  obtenido de la distribución  $\chi^2$ . En estos casos, se espera que los puntos  
 514 en el gráfico T2Qv generalicen el comportamiento de estas diferencias y superen el límite de control  
 515 superior ( $UCL$ ). La ubicación de este límite de control varía en función del número de dimensiones  
 516 que se representen, así, cuando es alto se logra un desempeño óptimo, mientras que, se introduce  
 517 inestabilidad y se pierde confiabilidad en los resultados al disminuir el número de dimensiones de  
 518 entre las que se puede representar.

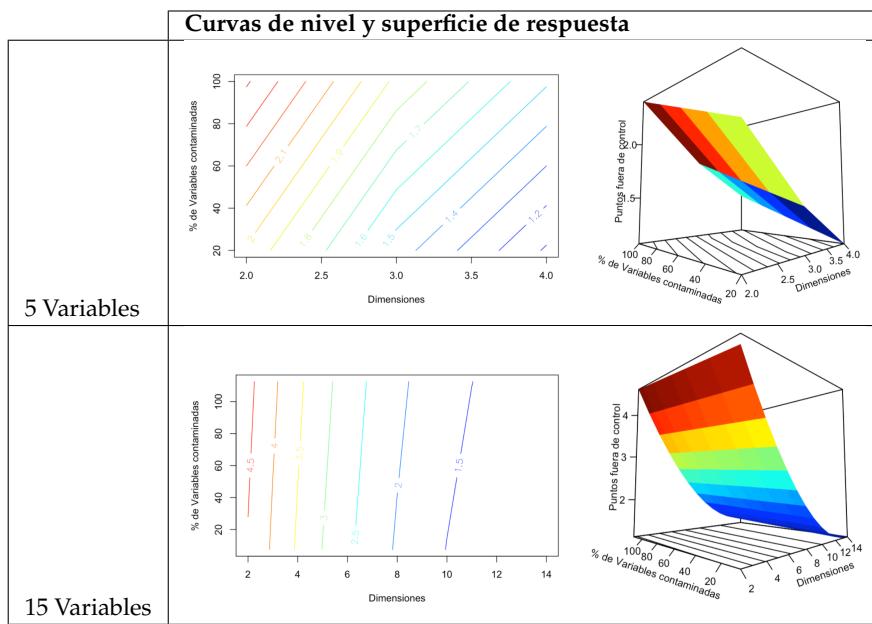
519 El gráfico de control propuesto es capaz de detectar un punto fuera de control, aún con un  
 520 bajo número de variables contaminadas, cuando se trabaja con un alto número de dimensiones. Se  
 521 recomienda  $p - 1$ , tal que  $p$  es el número total de dimensiones de la matriz inicial (Tabla 3). Cuando se  
 522 disminuye el número de dimensiones también disminuye la altura del límite de control superior ( $UCL$ ),  
 523 en consecuencia, se incrementa el número de puntos fuera de control, aunque no necesariamente las  
 524 variables expresen diferencias significativas en su valores, crece la probabilidad de falsos positivos.  
 525 Por consiguiente, la pregunta que surge es hasta cuántas dimensiones se puede disminuir en el análisis  
 526 sin perder confiabilidad en el resultado. La importancia de esta pregunta radica en la necesidad de  
 527 disponer un gráfico confiable, que identifique puntos fuera de control aún si se ha aplicado a los datos  
 528 una técnica de una reducción de dimensiones, sin caer en casos de falso positivo.



**Figure 14.** Curvas de nivel y superficie de respuesta obtenidas con el gráfico T2 Hotelling para variables cualitativas.

529 El análisis de sensibilidad utiliza curvas de nivel y superficies de respuesta (figura 14)  
 530 para representar el número de puntos fuera de control, considerando el porcentaje de variables  
 531 contaminadas de la  $k_i$  tabla y el número de dimensiones representadas. Los datos de prueba utilizados  
 532 en el modelo se registran en 10 tablas, cada una de ellas incluye 10 variables y cada variable tiene tres  
 533 categorías: alto, medio y bajo. La tabla 10 tiene una distribución diferente de las demás, esta es la tabla  
 534 contaminada.  
 535 Se observa que el modelo es capaz de identificar un punto fuera de control trabajando con 9  
 536 dimensiones ( $p-1$ ), aún con un porcentaje bajo de variables contaminadas. Cuando el número de  
 537 dimensiones disminuye a 8 y el porcentaje de variables contaminadas es cercano a 100%, detecta  
 538 correctamente 1 punto fuera de control. Se observa además que cuando el número de dimensiones  
 539 es menor se pierde estabilidad. En consecuencia, el análisis de sensibilidad ratifica que el gráfico de  
 540 control T2Qv tiene un buen rendimiento cuando trabaja con altas dimensiones.

541 Con la finalidad de conocer el comportamiento del gráfico T2Qv con distintos números de  
 542 variables, se presenta 2 casos (Tabla 12), además del que se observa en la figura 14, que consta de 10  
 543 variables. Ambos casos presentan inestabilidad con dimensiones bajas. El primer caso está realizado  
 544 con 5 variables, 10 tablas, donde la última tiene una distribución diferente a las demás, a diferencia del  
 545 caso con 10 variables, este no es muy estable aún cuando el número de dimensiones se acerca al número  
 546 de variables. El segundo caso presenta la misma estructura de tablas pero consta de 15 variables. Se  
 547 denota que con mayor número de dimensiones es estable, aún si se disminuye 2 dimensiones sigue  
 548 detectando un solo punto fuera de control, que corresponde al escenario correcto. De este modo se  
 549 comprueba que el gráfico T2Qv es más estable con mayor cantidad de variables, sin embargo, los  
 550 resultados con pocas variables siguen siendo fiables.



**Table 12.** Curvas de nivel y superficie de respuesta obtenidas con el gráfico T2 Hotelling para variables cualitativas con 5 y 15 variables.

## 551 6. Discusión

552 En el SPC para variables cualitativas todavía no son muchas las propuestas publicadas. Las  
 553 diferencias entre procedimientos para la determinación de los estadísticos y los gráficos de control en  
 554 este campo hacen difícil su comparación.

555 El gráfico de control T2Qv, que se presenta en este artículo, aplica un MCA, técnica de análisis  
 556 multivariante que identifica estructuras latentes que subyacen en el conjunto de datos cualitativos  
 557 y que involucra una reducción de dimensiones, en consecuencia, desde el comienzo se requiere una  
 558 tabla de datos con  $p$  variables ( $p > 3$ ) dicotómicas o polítómicas. Se debe recordar que el análisis  
 559 de sensibilidad determinó que esta propuesta tiene un buen rendimiento cuando trabaja con altas  
 560 dimensiones y que a bajas dimensiones pierde estabilidad. En varios estudios revisados, los casos  
 561 de aplicación analizan sólo dos o tres variables, lo que conduciría a la aplicación de un análisis de  
 562 correspondencias simple, no múltiple. En consecuencia, estos casos no podrían ser tratados con el  
 563 T2Qv.

564 Como ejemplos se señala la Combinación lineal óptima de variables Poisson para el SPC  
 565 multivariados, de Epprecht *et al.* [26] cuyo caso de aplicación registrado en su publicación analiza dos  
 566 variables relacionadas con el conteo de defectos en la producción de jarrones de cerámica. El gráfico  
 567 GMDS de Ali and Aslam [32] fue exemplificado con un conjunto de datos de telecomunicaciones,  
 568 tomado de Jiang *et al.* [45], que consta de sólo dos variables. El gráfico de control multivariante,  
 569 desarrollado por Fernández *et al.* [31]), para  $p$  características de calidad de atributos correlacionadas,  
 570 que aplica teoría difusa, hace un análisis de dos tablas tomadas de publicaciones de Taleb [30] y Taleb  
 571 *et al.* [29]), la primera con tres variables relacionadas con la comida congelada, y la segunda, con tres  
 572 variables sobre la producción de porcelana.

573 Otra de las características del gráfico T2Qv es que cada muestra es un grupo constituido por un  
 574 conjunto de individuos. El ejemplo de datos simulados *Datak10Contaminated* incluye un conjunto de  
 575 10 tablas y 11 variables, cada tabla es una muestra, está formada por 100 observaciones y aparece  
 576 representada como un punto en el gráfico  $T^2$  de Hotelling; el ejemplo aplicado al contexto educativo  
 577 hace referencia a la base de datos *Estudiantes 2019\_2020*, conformada por 43191 observaciones y 17  
 578 variables cualitativas agrupadas en 4 períodos académicos, estos períodos constituyen las tablas  
 579 (muestras) que se representan como puntos en el gráfico. En publicaciones de varios autores se puede

580 constatar que en sus ejemplos de aplicación se analiza una sola tabla, de dimensiones  $n$  (filas)  $\times p$  (variables), donde cada  $n_i$  fila es una muestra.

581 Por ejemplo, el gráfico de control MNP, de Lu [27] contiene en su artículo una tabla de datos simulados  
582 de 30 muestras, donde cada una de ellas es un único individuo (objeto) que registra el conteo de  
583 defectos para tres características de la calidad. Asimismo, la exemplificación que Chiu and Kuo [23]  
584 presentaron de su gráfico de control MP se hizo con una tabla de datos simulados de 26 muestras,  
585 donde cada muestra representa a un individuo al que se le registra el  $D$  número de defectos o no  
586 conformidades asociadas a tres características de calidad.

587 En el gráfico de control T2Qv que se presenta en este artículo, cada uno de los individuos (filas) que  
588 conforman las diferentes muestras pueden tener distintas configuraciones en función de las categorías  
589 de las variables. En base de datos *Estudiantes\_2019\_2020*, por ejemplo, el primer individuo de la  
590 lista es una mujer que estudia la carrera de Acuicultura en la Facultad de Ciencias Agropecuarias,  
591 su edad está entre 18 y 30 años, no presenta discapacidad, se autodeclaró mestiza; vive en una zona  
592 urbana, su padre tiene un nivel de formación de Educación Básica, su madre también; en su hogar  
593 viven 5 o más personas, sus estudios secundarios los realizó en un colegio particular, el ingreso total  
594 de su hogar se clasifica como de Rango 2, no registró el origen de los recursos económicos para sus  
595 estudios, no tuvo necesidad de acudir a segunda ni tercera matrícula y sí terminó su periodo académico.  
596 Otros estudiantes de esta misma tabla, o de las otras tres, tendrán diferentes características, hay que  
597 considerar que en total son 43191 individuos.

598 Por el contrario, otros autores que han investigado sobre gráficos de control multivariante para  
599 datos de atributos, aunque en su análisis consideran varias características de calidad, al final clasifican  
600 a cada individuo por una sola de las variables analizadas. Es el caso de Mukhopadhyay [28], cuya  
601 propuesta se demuestra con un caso de aplicación que controla 7 características de calidad en 24  
602 muestras cuyo tamaño varía entre 20 y 404 individuos. Las variables responden a 6 tipos de defectos  
603 de la pintura en la cubierta de ventiladores de techo: cobertura deficiente, desbordamiento, defecto de  
604 empanada, burbujas, defectos de pintura, defectos de pulido. La séptima característica es la ausencia  
605 de defectos. A cada individuo se lo clasifica por su defecto más predominante, por consiguiente, en  
606 su registro sólo aparece un tipo de defecto o ausencia de defectos, lo que resulta en una pérdida de  
607 información sobre el efecto combinado de las variables sobre el proceso.

## 608 7. Conclusiones

609 En este artículo se ha presentado el gráfico de control T2Qv, un técnica de control estadístico  
610 de procesos multivariantes que realiza un análisis de los datos cualitativos a través del Análisis de  
611 correspondencias múltiple, cuyas coordenadas se someten a un proceso de normalización propio del  
612 Análisis Factorial Múltiple, para luego representarlos mediante el gráfico  $T^2$  de Hotelling.

613 Esta propuesta genera un gráfico del MCA de la tabla consenso, que sirve de referente para comparar  
614 otros gráficos del MCA de las tablas que hayan sido identificadas como puntos fuera de control en  
615 el gráfico de Hotelling. Allí se puede verificar qué categorías de las variables han tenido variaciones  
616 en su ubicación en ambos gráficos, que pueden estar provocando cambios en la media del proceso y  
617 ocasionando el estado de fuera de control.

618 Para facilitar la interpretación del comportamiento de las variables se realiza un análisis de la distancia  
619 Chi cuadrado entre las categorías de la tabla consenso y de las tablas reportadas como fuera de control.  
620 Para este análisis se puede utilizar una tabla que reporta los valores del estadístico Chi cuadrado y  
621 los p-valores que determinan significancia estadística en tres niveles: 0.05 (\*), 0.01(\*\*) y 0.001(\* \*\*).  
622 También se puede representar este análisis mediante un gráfico de barras que incluye límites asociados  
623 a los niveles de significancia estadística establecidos.

624 El análisis de sensibilidad determinó que el gráfico de control T2Qv tiene un buen rendimiento cuando  
625 trabaja con altas dimensiones, pero, que pierde estabilidad a bajas dimensiones. Para facilitar la  
626 difusión y aplicación del método propuesto, se ha desarrollado un paquete estadístico computacional  
627 reproducible en R, denominado T2Qv y disponible en GitHub, que permite visualizar los resultados de

629 forma plana o interactiva, además, presenta un panel Shiny que contiene todas las funciones integradas  
630 en un mismo espacio.

631 En el SPC para variables cualitativas todavía no son muchas las propuestas publicadas. Las  
632 diferencias entre procedimientos para la determinación de los estadísticos y los gráficos de control en  
633 este campo hacen difícil su comparación.

634 **Appendix A**

635 *Appendix A.1*

636 **Appendix B**

637 **References**

- 638 1. Gutiérrez, H.; de la Vara Salazar, R. *Control estadístico de la calidad y seis sigma*; Vol. 3, McGraw Hill Education,  
639 2013; p. 152 – 253.
- 640 2. Ramos, M. Una alternativa a los métodos clásicos de control de procesos basada en coordenadas paralelas,  
641 métodos Biplot y Statis. PhD thesis, 2017.
- 642 3. Li, J.; Tsung, F.; Zou, C. Directional control schemes for multivariate categorical processes. *Journal of  
643 Quality Technology* **2012**, *44*, 136–154.
- 644 4. Hotelling, H. Multivariate quality control. Techniques of statistical analysis. *McGraw-Hill, New York* **1947**.
- 645 5. Lowry, C.A.; Woodall, W.H.; Champ, C.W.; Rigdon, S.E. A multivariate exponentially weighted moving  
646 average control chart. *Technometrics* **1992**, *34*, 46–53.
- 647 6. Crosier, R.B. Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics*  
648 **1988**, *30*, 291–303.
- 649 7. APARISI, F. Hotelling's T2 control chart with adaptive sample sizes. *International  
650 Journal of Production Research* **1996**, *34*, 2853–2862, [<https://doi.org/10.1080/00207549608905062>].  
651 doi:10.1080/00207549608905062.
- 652 8. Aparisi, F.; Haro, C.L. Hotelling's T2 control chart with variable sampling intervals. *International  
653 Journal of Production Research* **2001**, *39*, 3127–3140, [<https://doi.org/10.1080/00207540110054597>].  
654 doi:10.1080/00207540110054597.
- 655 9. Faraz, A.; Parsian, A. Hotelling's T2 control chart with double warning lines. *Statistical Papers* **2006**,  
656 *47*, 569–593. doi:10.1007/s00362-006-0307-x.
- 657 10. Shabbak, A.; Midi, H. An improvement of the hotelling statistic in monitoring multivariate quality  
658 characteristics. *Mathematical Problems in Engineering* **2012**, 2012.
- 659 11. Kim, S.B.; Jitpitaklert, W.; Park, S.K.; Hwang, S.J. Data mining model-based control charts for multivariate  
660 and autocorrelated processes. *Expert Systems with Applications* **2012**, *39*, 2073–2081.
- 661 12. Ruiz-Barzola, O. Gráficos de Control de Calidad Multivariantes con Dimensión Variable. PhD thesis,  
662 Universitat Politècnica de Valéncia, 2013.
- 663 13. Yeong, W.C.; Khoo, M.B.C.; Teoh, W.L.; Castagliola, P. A control chart for the multivariate coefficient of  
664 variation. *Quality and Reliability Engineering International* **2016**, *32*, 1213–1225.
- 665 14. Gower distance-based multivariate control charts for a mixture of continuous and categorical variables.  
666 *Expert Systems with Applications* **2014**, *41*, 1701–1707. doi:10.1016/j.eswa.2013.08.068.
- 667 15. Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Multivariate control chart based on PCA  
668 mix for variable and attribute quality characteristics. *Production & Manufacturing Research* **2018**, *6*, 364–384,  
669 [<https://doi.org/10.1080/21693277.2018.1517055>]. doi:10.1080/21693277.2018.1517055.
- 670 16. Liu, Y.; Liu, Y.; Jung, U. Nonparametric multivariate control chart based on density-sensitive novelty  
671 weight for non-normal processes. *Quality Technology & Quantitative Management* **2020**, *17*, 203–215.
- 672 17. YILMAZ, H.; Yanik, S. Design of Demerit Control Charts with Fuzzy c-Means Clustering and an  
673 Application in Textile Sector. *Textile and Apparel* **2020**, *30*, 117–125.
- 674 18. Farokhnia, M.; Niaki, S.T.A. Principal component analysis-based control charts using support  
675 vector machines for multivariate non-normal distributions. *Communications in Statistics -*

- 676        *Simulation and Computation* **2020**, *49*, 1815–1838, [<https://doi.org/10.1080/03610918.2018.1506032>].  
677        doi:10.1080/03610918.2018.1506032.
- 678        19. Xue, L.; Qiu, P. A nonparametric CUSUM chart for monitoring multivariate serially correlated processes.  
679        *Journal of Quality Technology* **2020**, pp. 1–14.
- 680        20. Ahsan, M.; Mashuri, M.; Wibawati.; Khusna, H.; Lee, M.H. Multivariate Control Chart Based on  
681        Kernel PCA for Monitoring Mixed Variable and Attribute Quality Characteristics. *Symmetry* **2020**, *12*.  
682        doi:10.3390/sym12111838.
- 683        21. Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Outlier detection using PCA  
684        mix based T2 control chart for continuous and categorical data. *Communications in Statistics  
685        - Simulation and Computation* **2021**, *50*, 1496–1523, [<https://doi.org/10.1080/03610918.2019.1586921>].  
686        doi:10.1080/03610918.2019.1586921.
- 687        22. Holgate, P. Estimation for the bivariate Poisson distribution. *Biometrika* **1964**, *51*, 241–287.
- 688        23. Chiu, J.E.; Kuo, T.I. Attribute control chart for multivariate Poisson distribution. *Communications in  
689        Statistics-Theory and Methods* **2007**, *37*, 146–158.
- 690        24. Lee, L.H.; Costa, A.F.B. Control charts for individual observations of a bivariate Poisson process. *The  
691        International Journal of Advanced Manufacturing Technology* **2009**, *43*, 744–755.
- 692        25. Laungrungrong, B.; M, C.B.; Montgomery, D.C. EWMA control charts for multivariate Poisson-distributed  
693        data. *International Journal of Quality Engineering and Technology* **2011**, *2*, 185–211.
- 694        26. Epprecht, E.K.; Aparisi, F.; García-Bustos, S. Optimal linear combination of Poisson variables for  
695        multivariate statistical process control. *Computers & operations research* **2013**, *40*, 3021–3032.
- 696        27. Lu, X. Control chart for multivariate attribute processes. *International Journal of Production Research* **1998**,  
697        *36*, 3477–3489.
- 698        28. Mukhopadhyay, A.R. Multivariate attribute control chart using Mahalanobis D 2 statistic. *Journal of Applied  
699        Statistics* **2008**, *35*, 421–429.
- 700        29. Taleb, H.; Limam, M.; Hirota, K. Multivariate fuzzy multinomial control charts. *Quality Technology &  
701        Quantitative Management* **2006**, *3*, 437–453.
- 702        30. Taleb, H. Control charts applications for multivariate attribute processes. *Computers & Industrial Engineering*  
703        **2009**, *56*, 399–410.
- 704        31. Fernández, M.N.P.; García, A.C.; Barzola, O.R. Multivariate multinomial T 2 control chart using fuzzy  
705        approach. *International Journal of Production Research* **2015**, *53*, 2225–2238.
- 706        32. Ali, M.R.; Aslam, M. Design of control charts for multivariate Poisson distribution using generalized  
707        multiple dependent state sampling. *Quality Technology & Quantitative Management* **2019**, *16*, 629–650.
- 708        33. López, C.P. *Técnicas de análisis multivariante de datos*; Pearson Educación, 2004.
- 709        34. Pearson, K. LIII.On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series  
710        6* **1901**, *2*, 417 – 441. doi:10.1080/14786440109462720.
- 711        35. Hotelling, H. Analysis of a complex of statistical variables into principal components. **1933**. *24*, 417 – 441.  
712        doi:10.1037/h0071325.
- 713        36. Ch, S.; others. General intelligence objectively determined and measured. *American Journal of Psychology*  
714        **1904**, *15*, 201–293.
- 715        37. Thurstone, L.L. Multiple-factor analysis; a development and expansion of The Vectors of Mind. **1947**.
- 716        38. Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200.
- 717        39. Benzécri., J. *OL'analyse des correspondances. En L'Analyse des Données: Leçons sur L'analyse Factorielle et la  
718        Reconnaissance des Formes et Travaux*; Paris - 1973, 1973.
- 719        40. Michailidis, G.; Leeuw, J.D. The Gifi system of descriptive multivariate analysis. *Statistical Science* **1998**, pp.  
720        307–336.
- 721        41. Gifi, A. *Nonlinear multivariate analysis*; Vol. 14, John Wiley & Sons, 1990.
- 722        42. López de Ipiña, F. Análisis multivariante aplicado al estudio del parentesco. Representaciones HOMALS.  
723        **2014**.
- 724        43. Escofier, B.; Pagès, J. Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis*  
725        **1994**, *18*, 121–140. doi:[https://doi.org/10.1016/0167-9473\(94\)90135-X](https://doi.org/10.1016/0167-9473(94)90135-X).
- 726        44. Montgomery, D.C. *Statistical quality control*; Wiley Global Education, 2012.
- 727        45. Jiang, W.; Au, S.; Tsui, K.L.; Xie, M. Process monitoring with univariate and multivariate c-charts. *Technical  
728        Report, the Logistics Institute, Georgia Tech, and the Logistics Institute-Asia Pacific* **2002**.

729 © 2021 by the authors. Submitted to *Water* for possible open access publication under the terms and conditions of  
730 the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).